DOCUMENT RESUME

ED 258 993 TM 850 361

AUTHOR	Kuipers, Benjamin
TITLE	Expert Causal Reasoning and Explanation.
SPONS AGENCY	National Library of Medicine (DHHS/NIH), Bethesda, Md.; National Science Foundation, Washington, D.C.
PUB DATE	Mar 85
GRANT	NIH-LM-03603; NIH-LM04125; NIH-RO1-LM-04374; NSF-DCR-8417934; NSF-MCS-8303640
NOTE	27p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Artificial Intelligence; *Clinical Diagnosis; Cognitive Processes; Computer Simulation; Educational Psychology; Epistemology; *Inferences; Problem Solving; Research Methodology
IDENTIFIERS	*Causal Reasoning; Cognitive Psychology; Knowledge Representation; *Protocol Analysis

ABSTRACT

The relationship between cognitive psychologists and researchers in artificial intelligence carries substantial benefits for both. An ongoing investigation in causal reasoning in medical problem solving systems illustrates this interaction. This paper traces a dialectic of sorts in which three different types of causal resaoning for medical problem solving are identified in verbatim protocols and simulated by computer program. The first, and most commonly discussed, type of causal reasoning consists of "causal links" holding between states of the world. The second is based on "qualitative simulation" of systems of continuous parameters related by qualitative constraints. The third, the "one parameter simulation", is a hybrid of the first two. These three types of causal descriptions are not alternative hypotheses, but apparently coexist in the expert's mind. (BS)

Expert Causal Reasoning and Explanation

Benjamin Kuipers[1]
Clinical Decision Making Group,
MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139

March 28, 1985

Paper presented to the

American Educational Research Association
1985 Annual Meeting
Chicago, Illinois
March 31, 1985.

[1] This research was supported in part by the National Library of Medicine through NIH Grants LM 03603, LM 04125, and R01 LM 04374, and by the National Science Foundation through grants MCS-8303640 and DCR-8417034.

1 Introduction

The relationship between cognitive psychologists and researchers in artificial intelligence, although occasionally touchy, carries substantial benefits for both. The designer of expert systems frequently starts with a naive model of the type of reasoning he is attempting to implement. Careful review of the psychological literature and analysis of verbatim protocols can reveal unexpected properties of the reasoning he is investigating. In my experience studying the reasoning methods of expert physicians, careful attention to human behavior can reveal distinctions between radically different types of knowledge in what initially appeared to be a single category. The distinct types of knowledge need distinct representations in the design of an implementable system. In addition, clues to the actual structure of the knowledge representation appear in the form of states of partial knowledge.

The cognitive psychologist can learn from the AI researcher a vocabulary of knowledge representation and inference techniques that were developed for purely engineering purposes, but can serve as elements of descriptive psychological theories. The concepts of forward and backward-chaining inference rules provide examples of this. Implementation of a cognitive theory as a computer program also provides the well-recognized advantages of enforcing a certain level of consistency and completeness, and yields a computer program whose behavior can be considered a prediction, if treated very carefully. We will look at some examples of these interactions in the context of knowledge representations for reasoning about causal relations, specifically in medical problem-solving systems.

Causal reasoning is a phenomenon that has attracted much attention recently in both the cognitive science community [Gentner & Stevens, 1983] and in the artificial intelligence/expert systems community [de Kleer, 1977; de Kleer and Brown, 1984; Forbus, 1984; Kuipers, 1984, 1985]. Medical problem-solving systems such as MYCIN and Internist-I are fundamentally based on weighted associations between findings and hypotheses. In order to avoid a combinatorial explosion, such systems typically include assumptions about the independence of these associations, which leaves them unable to handle non-trivial interactions between diseases. Causal reasoning is seen as a way to avoid some of the limitations of these systems by incorporating knowledge about how the mechanisms of the body work. A *causal model* of a disease process and its evolution over time provides additional constraints that allow incompatible combinations of hypotheses to be excluded, and may permit the combined effects of two diseases to be predicted.

In this paper we will trace a dialectic of sorts, in which different types of causal reasoning are identified in verbatim protocols and simulated by computer programs. The first type of causal description consists of "causal links" holding between states of the world. This type of causality is the most commonly discussed in the scientific literature in psychology, philosophy, and artificial intelligence. The second type of causal description is based on "qualitative simulation" of systems of continuous parameters related by qualitative constraints. This is a qualitative abstraction of differential equations as a description of a physical system. The third type of description, the "one-parameter simulation", is a hybrid of the first two types. We have recently identified examples of this third type of causal reasoning in protocols, and are now developing specifications for a computer simulation.

These three types of causal descriptions are not alternative hypotheses, but apparently coexist in the expert's mind. Open problems include how they relate with each other, and which problems are most adequately handled by which type of knowledge. I present this discussion as an example of an on-going investigation into causal reasoning combining the points of view of the cognitive scientist and the AI knowledge representation designer.

2 The Causal Link

The most common representation for causal reasoning is the *Causal Link* representation which consists of states linked by relationships labeled *Causes* or *Caused-By*. The states, strictly speaking, are descriptions of aspects of the patient's overall condition. This type of representation is useful as a completeness and coherency criterion on explanations of the patient state. Ideally, the complete description should consist of a network of states which are either caused by other states or are acceptable as primary causes. Similarly, causal links are useful for generating hypotheses by specifying the possible causes of the states currently believed true.

A fragment of a verbatim transcript illustrates reasoning using the causal link representation. The subject has been presented with a few observations about the patient, and is attempting to construct a coherent explanation for her condition. The underlined words are the key phrases corresponding to the subsequent analysis.

L014 A: Well, they say that there's,
L015 that's she's clearly dehydrated
L016 and with postural hypotension,
L017 and so I'd be wondering the reasons why.
L018 She apparently hasn't been eating well,
L019 and, and I'd be concerned of whether
L020 she's had any g.i. losses,
L021 any vomiting,
L022 or any diarrhea,
L023 or any other things
L024 to cause the significant volume depletion that she seems to have.

The conceptual content of this fragment can be analyzed naturally as references to a number of state-descriptions and causal ($\Longrightarrow$), equivalence ($\equiv$), and specialization ($\longrightarrow$) relations between the states.

dehydration ≡ *volume-depletion*	*L015, L024*
volume-depletion ⟹ *postural-hypotension*	*L015, L016*
anorexia ⟹ *volume-depletion*	*L018*
gi-losses ⟹ *volume-depletion*	*L020, L024*
gi-losses ⟶ *vomiting, diarrhea, etc.*	*L020-L023*

The intermediate states, like ***volume-depletion*** and ***gi-losses***, provide organization for the set of possible causes of observed findings.

2.1 AI Research on Causal Links

The classic program based on causal links among states is CASNET [Weiss, et al, 1978] which diagnoses glaucoma at an expert level of performance. In CASNET, individual pathophysiological states are confirmed based on clinical findings of various kinds linked to them with a specified strength of association. The states are connected to each other with causal links weighted according to the frequency with which that particular causal pathway holds. A disease process corresponds to a path through the (non-cyclic) network of states. The current state of the disease process in a particular patient corresponds to partial progress along that path. The degree of belief in a particular state combines the support of direct observations for that particular state with support propagated from causally related states.

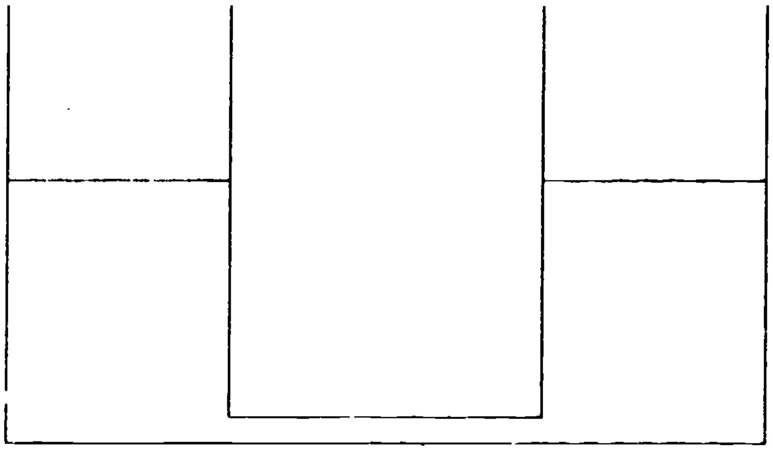
Rieger and Grinberg (1977) developed a taxonomy of different types of causal relations among nodes representing states, events, actions, and tendencies. Their representation enabled them to describe the behavior of mechanisms in terms of propagating activation of nodes in the causal network. The network could often be decomposed into loosely coupled modules, but was basically "flat", in that there was no explicit separation of different levels of description.

The ABEL system for acid-base and electrolyte diagnosis [Patil, 1981] provides multiple levels of state description, ranging from the clinical observation level down to a detailed description of physiological processes. The different levels allow clinical observations such as vomiting and dehydration to be mapped to physiological assertions such as decreased sodium or increased potassium concentrations in the blood. In ABEL, causal links are not weighted by frequency, but state descriptions have quantitative components representing the *amount* of the effect produced by the cause. This allows the system to reason about whether the observed causes are *sufficient* to account for the magnitude of the observed problem, or whether an additional factor should be sought. The ability to reason about the combination of reinforcing or compensating effects helps ABEL handle interactions among diseases.

Pople (1982) also incorporates a network of causal relations along with the taxonomic relations in his design for Cadeuceus.

2.2 U-Tube Example

Suppose we have two tanks of water connected by a pipe at the bottom of each. This is often described as a U-tube in elementary physics classes. The water in both tanks is at the same level, and there is no flow through the pipe.



If we suddenly add some water to tank A, there is flow through the connecting pipe until the system reaches a new equilibrium. A causal link description of this process would look something like the following.

$$\begin{aligned}
\uparrow level(A) &\Longrightarrow \uparrow pressure(A)\\
&\Longrightarrow \uparrow \Delta P\\
&\Longrightarrow \uparrow flow\\
&\Longrightarrow \downarrow level(A), \uparrow level(B)\\
&\Longrightarrow \downarrow pressure(A), \uparrow pressure(B)\\
&\Longrightarrow \downarrow \Delta P\\
&\Longrightarrow \downarrow flow\\
&\Longrightarrow flow = 0
\end{aligned}$$

Medical texts are full of similar diagrams consisting of terms connected with causal links, typically in a non-linear network. Our discussion is concerned with what formal representation these can correspond to.

2.3 Critique

There are several problems with the causal link representation that limit its ability to express relationships generally considered to be "causal" or to make certain pragmatically important causal inferences.

2.3.1 Semantics of the Causal Link

The causal chain presented above consists of terms of the form $\uparrow X$. Depending on the context of the particular diagram, such a term can be used to mean a variety of different things:

- X is increasing (i.e. has a positive derivative),

- the value of X is greater than normal,

- the value of X is greater than the previous value of X we considered,

- there is a tendency for X to increase which is combined with all other tendencies on X to determine its actual direction of change.

Similarly, on examining the causal links in the system, we see that similar links, of the form $\uparrow X \Longrightarrow \uparrow Y$, can be used to mean quite different things:

- $\downarrow flow \Longrightarrow flow = 0$ takes place over an interval in time,

- $\uparrow level(A) \Longrightarrow \uparrow pressure(A)$ takes place within the same instant in time.

The interpretation according to which causal relations take place over a temporal interval is certainly the most common, and is necessary to avoid contradiction in equilibrium situations like the above where $\uparrow level(A) \Longrightarrow \cdots \Longrightarrow \downarrow level(A)$. However, when compared with the physical situation, some causal relations can be seen to impose an ordering on events or changes that are physically constrained to take place simultaneously, as when $\uparrow temperature \Longrightarrow \uparrow pressure$ in a container of gas.

Treating these "causal" relationships as identical means that events that are actually simultaneous are treated as though they were spread over time. (For example, char acters on the Saturday morning cartoons frequently run off of cliffs, and yet have time for second thoughts before they start falling.) This distortion may be characteristic of at least some human reasoning, as it resembles the "Aristotelean physics" observed by McCloskey, et al (1980) in naive physics students. Thus, a psychologist interested in the cognitive development of causal reasoning might find such a collapse of distinct relations descriptively useful.

However, expert systems designers find such reasoning methods pragmatically undesirable, since they create intermediate state descriptions that are not physically realizable. It is thus difficult to validate the system's inferences or knowledge base against the scientific or technical literature in the expert domain.

The fact that the same terms are used to express importantly distinct types of assertions and relationships suggests that this type of causal description is less useful for predicting the behavior of an unknown mechanism than for explanation of predictions derived in some other fashion.

2.3.2 Local vs Global States

Properly speaking, causal relations should hold only between global states of the world. However, as they are typically used in medical texts or causal-link-type problem-solving systems, they hold between individual attributes. For example, in a model of the nephrotic syndrome we have studied elsewhere [Kuipers and Kassirer, 1984], we saw the relation:

$$\textit{decreased serum protein} \Longrightarrow \textit{increased interstitial fluid.}$$

This illustrates the point of the previous section, since the Starling equilibrium mechanism requires the patient to be simultaneously in the two states, *decreased serum protein* and *increased interstitial fluid*. The stated relationship is only sometimes true, and can be blocked if *decreased serum sodium* is also true. A physiological mechanism depends on a richly structured set of relationships among the different attributes. In order for a causal description of a mechanism to be useful for predicting future states, it must be able to express that complex set of relationships.

2.3.3 Predicting Behavior from Structure

In causal reasoning about physical mechanisms, a paradigmatic type of reasoning is predicting the behavior of a mechanism from the behavior of its parts and the relationships between them. Since the nodes of a causal network are states or events (i.e. fragments of potential behavior), there is no representation for the structure of a mechanism as distinct from its behavior. The actual behavior in response to a particular situation is some selection of nodes and links in the network. While the combinations of states that are activated under a particular set of circumstances may be novel, there is no natural way to express the discovery of previously unsuspected states or behavior, such as the existance of an equilibrium point between two landmark values.

3 Qualitative Simulation

In search of a representation for causal reasoning that would be more adequate to explain the ability of an expert to predict the behavior of a mechanism under unexpected circumstances, we [Kuipers and Kassirer, 1983, 1984] examined transcripts of causal explanations. We found evidence suggesting that the structure of a mechanism is represented separately from its behavior. This, along with a new line of AI research, led us to focus on qualitative simulation as an alternative to the causal link.

The qualitative simulation approach to causal reasoning separates the description of the ***structure*** of a mechanism from the description of its ***behavior***. The structure of a mechanism is described in terms of continuously-variable parameters and constraints among them. Behavior is described in terms of the ordinal relations among the values of parameters and limiting landmark values, and their directions of change. The semantics of this representation can be made precise by creating a correspondence between the structural and behavioral descriptions and differential equations and the functions that satisfy them (figure 1).

The advantage of this approach to causal reasoning is that predictions of behavior can be made from the structural description. It is capable of inferring unexpected types of behavior, can create new landmark values where significant qualitative changes take place, and can handle feedback phenomena.

The following protocol fragment (analyzed more completely in Kuipers and Kassirer (1984)) illustrates the difference between time-independent facts about the structural relationships between values of two parameters (L162-178), and time-dependent behavioral facts about the events at some particular moment (L179-181).

$$\begin{tikzcd}
\text{Physical System} \arrow[r] \arrow[d] & \text{Actual Behavior} \arrow[d] \\
\text{Differential Equation} \arrow[r, "\text{numerical or analytic solution}"] \arrow[d] & f_i : \Re \rightarrow \Re \arrow[d] \\
\text{Structural Description} \arrow[r, "\text{qualitative simulation}"] & \text{Behavioral Description}
\end{tikzcd}$$

Figure 1: Qualitative simulation and differential equations are both abstractions of actual behavior.

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.

The analysis can be described as follows.

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, blo d → 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*

The detailed analysis demonstrates that there is a distinction in the representation between structural and behavioral descriptions of a mechanism. The explanation focuses on the relationships among and changes of continuous-valued parameters. And the values of those parameters are described in qualitative terms. As a psychological theory, of course, each of these conclusions is quite tentative, and is subject to further experimental exploration and evaluation. Nonetheless, they served as a valuable inspiration to a new and useful knowledge representation.

3.1 AI Research on Qualitative Simulation

There has been a recent surge of interest in qualitative simulation in AI, with theories by de Kleer, Forbus, Kuipers, Williams and others reported in a recent special issue of the *Artificial Intelligence Journal* (1984). The structure of a mechanism is described qualitatively in terms of a set of continuously varying quantities, linked by constraints representing the structural relations of the mechanism. Some constraints specify familiar mathematical relationships: $DERIV(vel, acc)$, $ADD(net, out.in)$, $MULT(mass, acc, force)$, $MINUS(fwd, rev)$. Others assert qualitatively that there is a functional relationship between two physical parameters, but only specify that the relationship is monotonically increasing or decreasing: $M^+(price, power)$ and $M^-(mph, mpg)$. The value of a parameter at any point in time is describe qualitatively in terms of its ordinal relations with a set of landmark values, and its direction of change. The behavior of the mechanism is described as the sequence of qualitative states taken on by the parameters.

Differences among AI approaches to qualitative simulation include the form of the constraints, how the constraint sets are created, what landmarks are known, whether landmarks are linearly ordered, whether new landmarks can be created, and the algorithm used by the reasoning process.

In all cases, the qualitative simulation algorithm derives a set of possible behaviors from the description of the structure of the mechanism. Ideally, the structural description will be well enough selected so that simulation yields a single behavior which describes the actual behavior of the mechanism, though at a more abstract level than a real-valued function. These techniques perform well at predicting the behavior of equilibrium situations under a variety of perturbing effects [Kuipers, 1984]. In more complex situations such as continuing or dissipating oscillation, it is possible for qualitative simulation to predict impossible behaviors [Kuipers 1985].

In the context of a diagnostic problem-solver it seems that the role of such a qualitative simulation is to generate the possible consequences of a hypothesized primary disease. The diagnostic system can then test whether the observed facts correspond to some possible scenario for the disease. The mechanism description both generates predictions to be tested, and elaborates a more

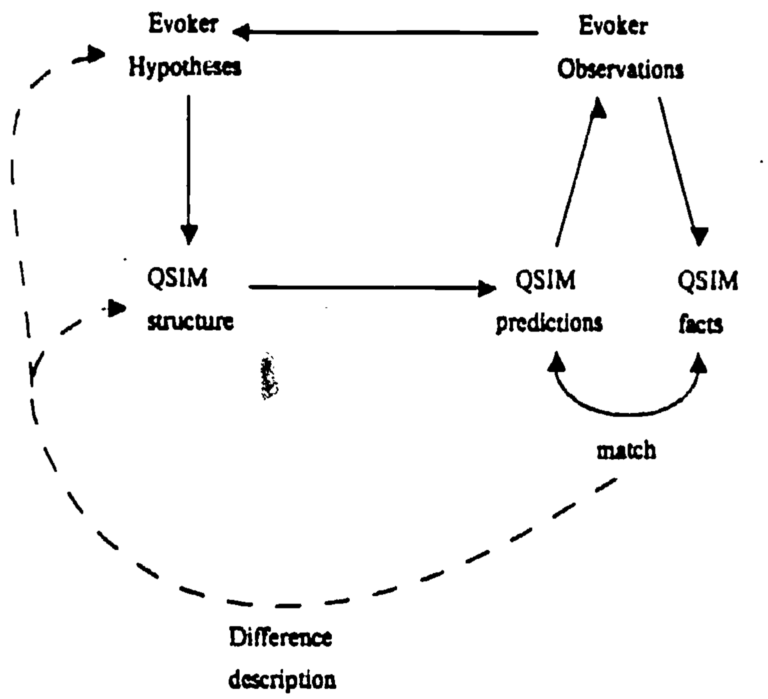


Figure 2: The Interaction between QSIM and the Evoker

detailed description of the patient state than would be possible otherwise. We are currently in the midst of creating a diagnostic program called RENAL which operates in this fashion (see figure 2), based on the interaction between a frame-based diagnostic program (the Evoker) and a qualitative simulation program (QSIM).

3.2 U-Tube Example

In our example of the U-tube, the structural description is stated in terms of continuous parameters for the level, pressure, pressure-difference, and flow across the pipe. (Figure 3)

The behavioral description is constructed by finding sets of qualitative transitions for the parameters in the structure, consistent with the constraints. Figure 4 shows the qualitative behavior as a sequence of qualitative states for each parameter. A qualitative state is a pair consisting of a magnitude and a direction of change (*increasing*, *decreasing*, or *steady*). A magnitude is either a landmark value (e.g. $PA*$ is the initial value of $pressure(A)$) or an open interval bounded by

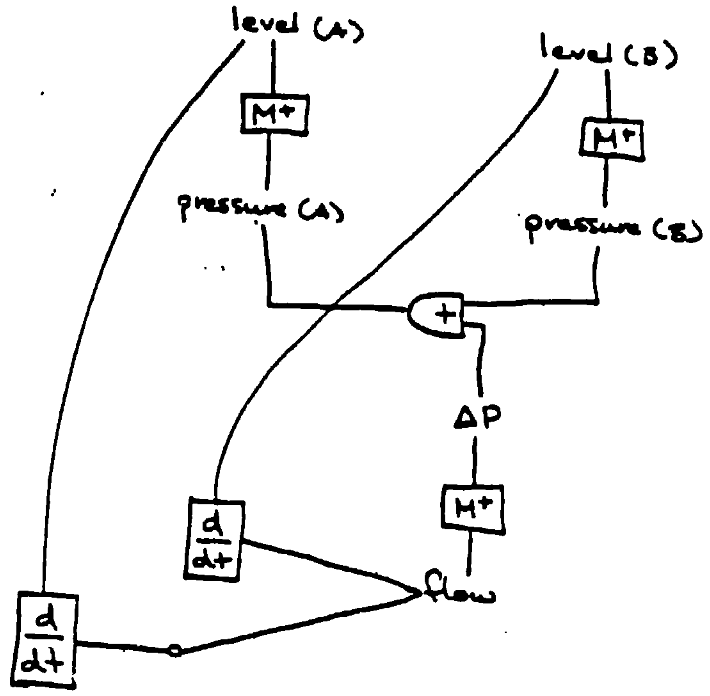


Figure 3: Structural Description of the U-Tube

landmark values (e.g. $(PA*, \infty)$). In this case, there is only one consistent behavior, so we know that it represents the actual behavior of the system [Kuipers 1985].

We can also describe the behavior as a set of *qualitative graphs* of the individual parameters, where only the *ordinal* relationships between the values on the axes and the points plotted are significant. (Figure 5)

3.3 Critique

Qualitative simulation predicts the behavior of a system correctly and uniquely when given a properly structured first-order description of the system (i.e. the corresponding differential equation includes only first derivatives). Modest branching takes place corresponding to genuine alternative behaviors consistent with the given structure and initial state. Most second-order systems and poorly constructed first-order systems yield widely branching behaviors that contain too many alternatives to be useful. This could be evidence that qualitative simulation is too complex to be

level(A)	$\langle (XA*, \infty), dec\rangle$	$P7 \rightarrow$	$\langle (XA*, \infty), dec\rangle$	$I9 \rightarrow$	$\langle XA1, std\rangle$
level(B)	$\langle XB*, inc\rangle$	$P5 \rightarrow$	$\langle (XB*, \infty), inc\rangle$	$I8 \rightarrow$	$\langle XB1, std\rangle$
pressure(A)	$\langle (PA*, \infty), dec\rangle$	$P7 \rightarrow$	$\langle (PA*, \infty), dec\rangle$	$I9 \rightarrow$	$\langle PA1, std\rangle$
pressure(B)	$\langle PB*, inc\rangle$	$P5 \rightarrow$	$\langle (PB*, \infty), inc\rangle$	$I8 \rightarrow$	$\langle PB1, std\rangle$
ΔP	$\langle (0, \infty), dec\rangle$	$P7 \rightarrow$	$\langle (0, \infty), dec\rangle$	$I5 \rightarrow$	$\langle 0, std\rangle$
flow	$\langle (0, \infty), dec\rangle$	$P7 \rightarrow$	$\langle (0, \infty), dec\rangle$	$I5 \rightarrow$	$\langle 0, std\rangle$

Figure 4: Behavior of the U-Tube: Parameter Transitions

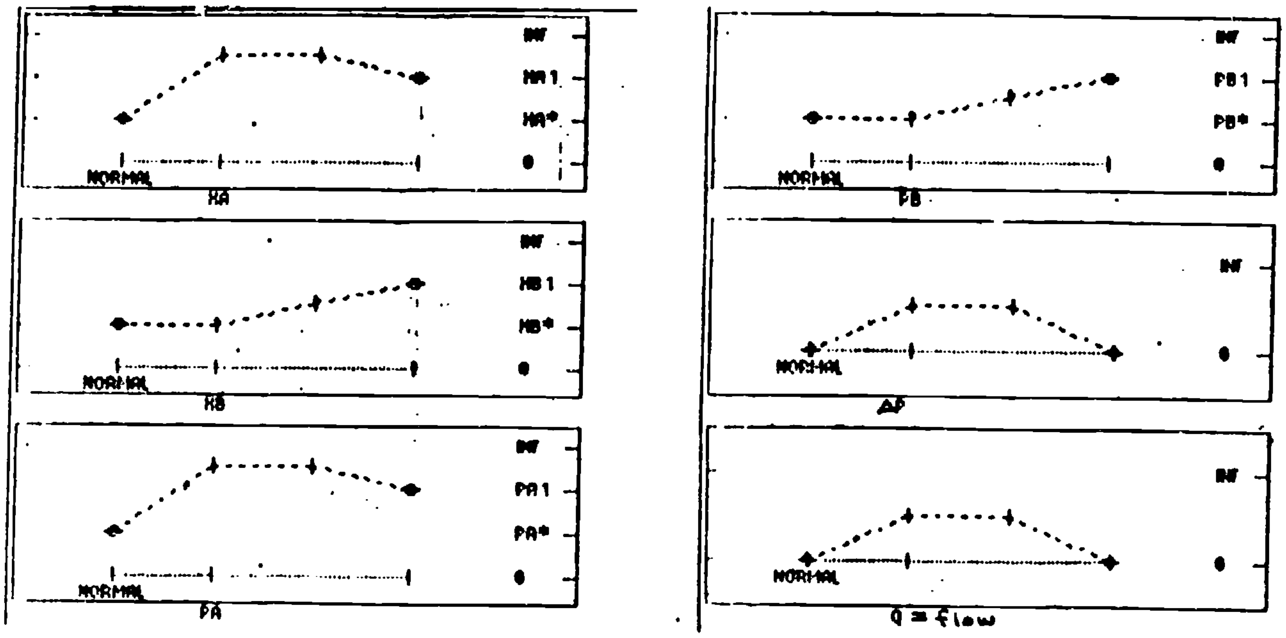


Figure 5: Behavior of the U-Tube: Qualitative Graphs

a realistic model of human causal reasoning. On the other hand, it is consistent with the plausible hypothesis that people can only use qualitative simulation on first-order models of mechanisms, and even then, only when their model of the structure has been debugged through training and experience.

Evidence from verbatim protocols suggested certain features of the knowledge representation that led to the development of qualitative simulation algorithms. However, the protocols do not contain clear references to all of the stages of those algorithms. There are several possible explanations, which require further empirical work to be distinguished.

- Only the initial propagation of information to create a complete initial state description is accessible to verbal explanation; the actual simulation takes place "automatically" and unconsciously.

- Propagation of the initial state description, simulation of subsequent states, and the description of the final state are all accessible to verbal explanation, but these very different computational processes are verbalized with similar constructions, making the correspondence difficult to determine.

- Qualitative simulation is only done at "learning time", and a generic behavior is retrieved to fit particular problems. An explanation selects which features of the stored information to verbalize, but has no correspondence to a trace of the computation.

- Qualitative simulation is a mathematical construct with a general resemblance to human causal reasoning, but does not, in fact, correspond to any cognitive process.

In the next section, I explore an alternative type of causal reasoning that combines some of the desirable features of the causal link model and the qualitative simulation model.

4 One-Parameter Simulation

Qualitative reasoning systems have generally been oriented toward complete descriptions of complex systems with subtle behavior. In such cases, a complex qualitative simulation algorithm is necessary in order to derive adequate results from the given problem statement. However, protocol analysis often reveals a simpler usage, focusing on the behavior of a single parameter as it moves to and past various landmark values.

Although the mechanism being simulated consists of a single parameter, and is thus much simpler than in the qualitative simulation case, the inference process may still involve sophisticated reasoning methods. In the fragment below, there are implicit references to a "health status" parameter that can be either stable or deteriorating, and to an unspecified future event, presumably when the patient reaches a point of no return, before which action is required.

L043 When I'm told that there is no improvement,
L044 and... when someone remains stable for forty eight hours,
L045 I think you're in a position
L046 where you can buy a little bit more time.
L047 If there's deterioration,
L048 which I'm not told,
L049 then I'd feel a little more strongly about moving ahead.

In order to capture the content of this fragment, the causal reasoner must be able to express alternate hypothetical worlds, the qualitative magnitude and direction of change of continuous parameters, the magnitudes of time-intervals as well as physical parameters, and the comparison of magnitudes across hypotheses.

Actual(Now)	*L044*
Health-Status(Patient,Now) = ⟨poor, stable⟩	*L043-L044*
Hypothetical(Now2)	*L048*
Health-Status(Patient,Now2) = ⟨poor, deteriorating⟩	*L047*
Event = last opportunity for treatment	*implicit*
Time-Interval(Now2,Event) < Time-Interval(Now,Event)	*L045-L046, L049*

This comparison between two hypothetical situations for a single patient is generalized and abstracted to a monotonic relationship between health status and urgency of treatment. In the following fragment from the same protocol, this consideration is tied to the selection of an invasive test.

L070 What I use to decide which to go to is really ...
L071 partly, how sick the patient is.
L072 'Cause I actually think
L073 the sicker a patient is the more rapidly the ————
L074 (the) more likely I am
L075 to go to an open lung biopsy.

M^+ (Health-Status(Patient,Now), Time-Interval(Now,Event))	*L072-l073*
M (Time-Interval(Now,Event), Preference(Open-Lung-Biopsy,Bronchoscopy))	*L074-L075*

In the one-parameter simulation, rather than using a network of constraints on a set of simultaneously changing parameters to determine which of many possible combinations of behaviors are consistent, simulation projects the possible futures of a single driving parameter. The remaining features of the current state description can then be inferred, if needed, from the qualitative value of the driving parameter. Prediction of the next state is done by moving the driving parameter in its direction of change and checking for the consistency of the resulting state. As we see in the above protocol fragment, a particularly useful application of this reasoning method might be to perform the same simulation in slightly varying contexts to determine an abstract relationship.

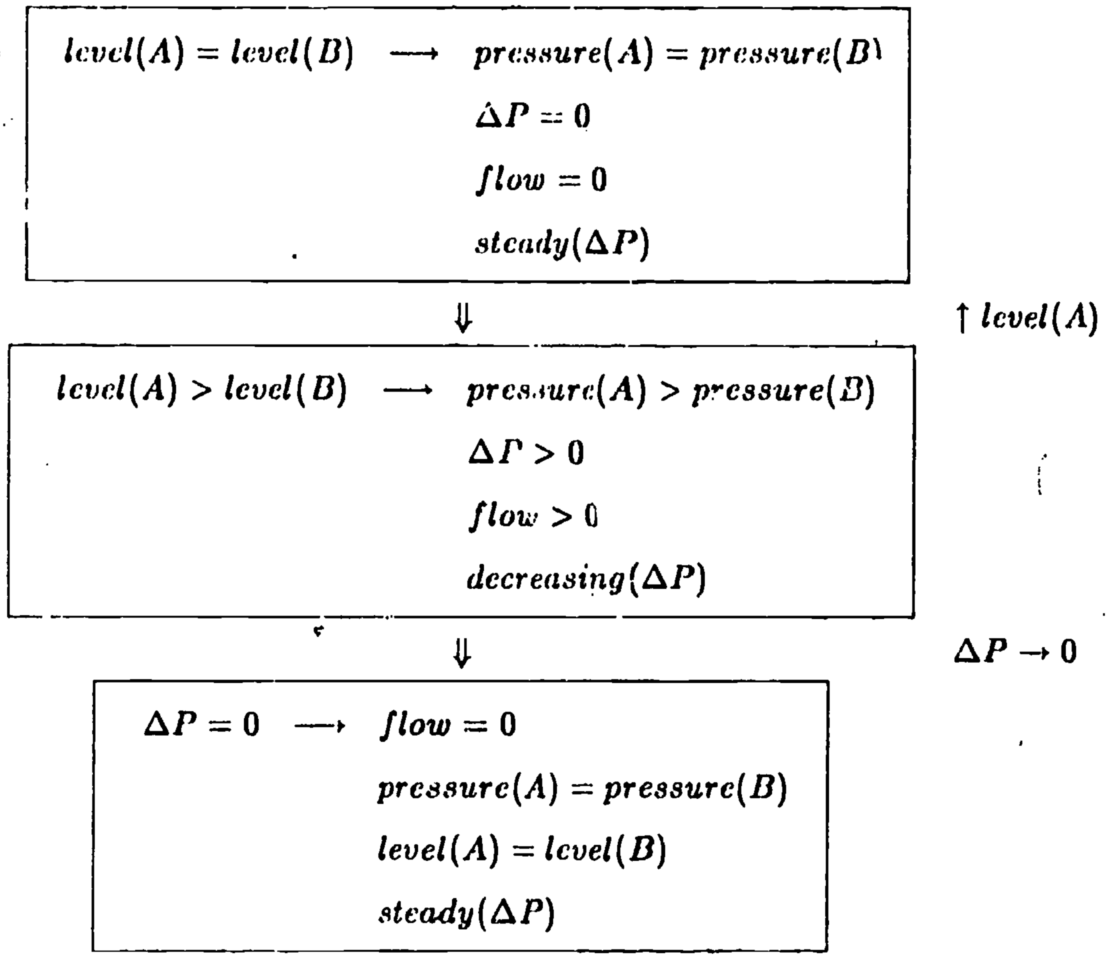

$level(A) = level(B) \longrightarrow pressure(A) = pressure(B)$
$\Delta P = 0$
$flow = 0$
$steady(\Delta P)$
$\Downarrow$
$\uparrow level(A)$
$level(A) > level(B) \longrightarrow pressure(A) > pressure(B)$
$\Delta P > 0$
$flow > 0$
$decreasing(\Delta P)$
$\Downarrow$
$\Delta P \to 0$
$\Delta P = 0 \longrightarrow flow = 0$
$pressure(A) = pressure(B)$
$level(A) = level(B)$
$steady(\Delta P)$

Figure 6: One-Parameter Simulation of tne U-Tube

4.1 U-Tube Example

In the one-parameter simulation, the U-tube has the same structural description as for qualitative simulation (figure 3), but the structure is used only for propagation to fill out the current state description. Prediction of the next state is done by focusing on a single parameter, changing it as desired, and propagating to fill out the state.

In Figure 6, the top box is the initial state description, derived from the assumption that $level(A) = level(B)$. The first causal link ($\Downarrow$) corresponds to adding water to tank A, so that *level*(A) is increased; the remaining aspects of the state description follow by propagation. The second link corresponds to the selection of ΔP as driving parameter, moving to the limiting value 0. Other selections of driving parameter would have produced the same result.

Since the prediction phase focuses on a single parameter and does not watch the simultaneous evolution of all parameters, it does not conclude whether *level*(A) in the final state is greater or less than *level*(A) in the initial state. This demonstrates a trade-off between the amount of information deduced, and the robustness and computational cost of the algorithm.

4.2 Critique

The one-parameter simulation is not as powerful as the full qualitative simulation, since it does not capture interactions among parameters. It is not straight-forward for the one-parameter simulation to conclude, for example, that a new equilibrium point exists between two landmark values previously believed to be adjacent.

However, because it focuses attention on a single changing parameter, it is computationally simpler. Furthermore, based on preliminary study, it does correspond well with parts of the verbatim protocol that have not matched the qualitative simulation algorithm. It also appears to match certain less formal observations of physicians' behavior and explanations collected by Harry Pople [personal communication, 1983] in a review of open problems inspired by his research on the Internist/Caduceus system.

Both the computational and the empirical implications of the one-parameter simulation require considerable further study. Based on the examples we have collected that suggest the existence and properties of one-parameter simulation, a more systematic analysis of protocols is needed to establish those properties more clearly. We have also designed and are beginning to implement a working version of the one-parameter simulation.

5 Discussion

These examples illustrate three distinct types of causal reasoning and explanation that can be identified in verbatim transcripts. The underlying conceptual frameworks for the causal link representation and the qualitative simulation representation are almost completely distinct. The third type, however, suggests that they represent the poles of a spectrum of representations, across which the advantages and disadvantages of the two approaches are combined in various proportions. Thus, when we look at human behavior and hope to determine *the* reasoning technique in a certain area, we are likely to find, as in these examples, that a variety of techniques are used opportunistically.

The methods we have used for collection and analysis of verbal protocols are much more thoroughly discussed elsewhere [Ericsson and Simon, 1980; Kassirer, Kuipers, and Gorry, 1982; Kuipers and Kassirer, 1984]. The examples used here are from "thinking aloud" interviews where physicians are presented with cases in small packets, and encouraged in non-directive ways to think aloud while analyzing the case. A fragment of protocol is likely to provide better insights into the problem-solving process if it represents a point where the subject is clearly in the midst of solving the problem. A summarized conclusion is often in such a conventional form as to hide any traces of the actual problem-solving process. Although we interview physicians at several levels of expertise, we have generally found it more fruitful to study those at a "journeyman" level of expertise (e.g. second or third-year residents), than the "masters" who can leap directly from the problem to a correct answer.

The analysis takes place in two stages. First, we find a underlying domain of conceptual objects corresponding to all of the referring phrases found in a protocol fragment. Second, we attempt to devise a knowledge representation and inference process corresponding to the nature and order of the assertions we see in the fragment. Both the referring phrase analysis and the assertional analysis are "analysis by synthesis" processes drawing heavily on the analyst's familiarity with a range of knowledge representations and formal inference strategies. Therefore, although this type of analysis is very fruitful in suggesting knowledge-representation structures and their properties, it is dependent both on the state of theoretical work on knowledge representations and on the individual analyst.

Although care can be taken to avoid many known methodological pitfalls, at the current state of the cognitive sciences, there is always the problem of the observer being unable to recognize a phenomenon which is not expressible in his or her conceptual vocabulary. The benefit to the cognitive scientist of familiarity with knowledge representation research in artificial intelligence is just that: it provides a larger vocabulary of concepts with which to look at the cognitive world.

I offer these observations as a designer of knowledge representations with a reading knowledge of cognitive psychology, feeling that we AI researchers can use direction from psychologists about the actual nature of the knowledge we are trying to express, and that psychologists can benefit from familiarity with the range of representational tools we have available.

6 References

Artificial Intelligence Journal, Special volume on Qualitative Reasoning about Physical Systems. *Artificial Intelligence* **24**: 1-491. Also published as D. G. Bobrow (Ed.), *Qualitative Reasoning about Physical Systems.* New York: North-Holland, 1984.

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

J. de Kleer and J. S. Brown. 1984. A qualitative physics based on confluences. *Artificial Intelligence*, **24**: 7-83.

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

K. D. Forbus. 1985. Qualitative process theory. *Artificial Intelligence*, **24**: 85-168.

D. Gentner and A. Stevens, (Eds.). 1983. *Mental Models.* Hillsdale, NJ: Erlbaum.

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

B. J. Kuipers. 1984. Commonsense reasoning about causality: deriving behavior from structure. *Artificial Intelligence*, **24**: 169-203.

B. J. Kuipers. 1985. Qualitative simulation of mechanisms. MIT Laboratory for Computer Science TM-274. Submitted for publication.

B. J. Kuipers and J. P. Kassirer. 1984. Causal reasoning in medicine: analysis of a protocol. *Cognitive Science* **8**: 363-385.

J. McCarthy and P. J. Hayes. 1969. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie (Eds.), *Machine Intelligence* **4**. Edinburgh: Edinburgh University Press, 1969.

M. McCloskey, A. Caramazza, & B. Green. 1980. Curvilinear motion in the absence of external forces: naive beliefs about the motion of objects. *Science* **210**: 1139-1141.

R. S. Patil. 1981. Causal representation of patient illness for electrolyte and acid-base diagnosis. Cambridge. MA: MIT Laboratory for Computer Science TR-267.

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.

C. Rieger and M. Grinberg. 1977. The declarative representation and procedural simulation of causality in physical mechanisms. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, Cambridge, MA.

S. M. Weiss, C. A. Kulikowski, S. Amarel, and A. Safir. 1978. A model-based method for computer-aided medical decision-making. *Artificial Intelligence* 11: 145-172.

6.1 Acknowledgements

The protocol analysis techniques and some of the analyses themselves have been done with my collaborator Jerome P. Kassirer of the New England Medical Center. Christopher Eliot is responsible for implementing the qualitative simulation algorithms with very comfortable user interfaces, and for many valuable technical suggestions. I am also grateful to Harry Pople, developer of the Internist/Cadeuceus system, for the conversations which inspired the one-parameter simulation.