MIT/LCS/TR-318

A CONSTRAINT REPRESENTATION AND EXPLANATION FACILITY

Irwin J. Asbell

This blank page was inserted to preserve pagination.

A Constraint Representation and Explanation Facility

for

Renal Physiology

by

Irwin Joseph Asbell, M.D.

June 1984

Laboratory for Computer Science

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Cambridge Massachusetts 02139

A Constraint Representation and Explanation Facility

for

Renal Physiology

by

Irwin Joseph Asbell

Revised version of the thesis submitted to the Department of Electrical Engineering and Computer Science on August 16, 1982 in partial fulfillment of the requirements for the Degree of Master of Science

Abstract

Current research in Artificial Intelligence has yielded computer programs which have the potential to augment the physician's ability to diagnose illness. The medical diagnosis programs of the first generation contain medical facts representing associations between diseases and findings. A most important step is the development of computer programs that have models of physiological processes and have the ability to derive physiological justifications of observed signs and symptoms.

In this thesis, a program which models the causal mechanisms underlying a subset of human physiology is unveiled. We shall begin with a discussion of the relevant AI techniques used: envisionment, qualitative reasoning, propagation of constraints and grey boxes. To tie together these methodologies, the concept of explanation boxes is introduced. The rest of the thesis is devoted to the presentation of an explanation box network which is able to represent the physiological mechanisms pertaining to three syndromes of renal physiology.

Thesis supervisor: Peter Szolovits
Associate Professor of Electrical Engineering and Computer Science

This research was supported (in part) by the National Institutes of Health Grant No. 1 P01 LM 03374-03 from the National Library of Medicine.

Acknowledgments

I would like to thank:

Dr. E. Llewellyn-Thomas, Associate Dean of Medicine of the University of Toronto Medical School, for graciously granting me a year's leave of absence from medical school to help me study towards my S.M. degree.

My family in Toronto for their continuous encouragement.

Prof. Peter Szolovits, Prof. Ramesh Patil and other friends and colleagues at M.I.T. If I begin to list them, I am bound to leave someone out.

Ken Forbus for giving me a copy of his version of the CONLAN constraint language, upon which the program to be presented here was built.

My dear friends Manny and Eugenia Sopas of Arlington, Mass., whose warmth and friendship I will always remember.

Preface

In this thesis, an important research problem within the domain of Artificial Intelligence in Medicine (AIM) is addressed. For the reader interested in obtaining general background knowledge in the area of AI in medicine, Szolovits provides an enlightening account[26]. Here we shall focus on the issue of developing appropriate computer models for physiological systems. The computer program to be unveiled here, NEPHROS, models three syndromes of renal physiology.

Since the concepts to be presented encompass both computer science and medicine, I have made every effort to ensure that the material be readable to a wide audience. For the computer scientist, relevant medical concepts are introduced where necessary. In addition, a glossary of medical terms used is provided as an appendix. The first occurrence of all words defined in the glossary is underlined in the text. For those solely interested in medical applications, the material has been appropriately segmented to enable the reader to tailor his or her reading.

The computer program uses a knowledge base of medical information. For the purpose of keeping this information up-to-date within reason, the latest version of *Harrison's Principles of Internal Medicine* has been used[13]. Since AI in medicine programs must keep up with the fast pace at which medicine changes, they are now constructed with the ability to acquire new information and remove that which has become obsolete. As the presentation unfolds, it will be evident that NEPHROS has been designed with considerable modularity and hence facilitates the update task.

I hope the reader will find this thesis interesting and finishes with an enthusiasm for the development of solutions to the research problem discussed here.

I. A.

August, 1982
Cambridge, Mass.

CONTENTS

FIGURES

1. Introduction

1.1 Physiology and Artificial Intelligence

Guyton defines human physiology as an attempt at explaining "the specific characteristics and mechanisms of the human body that make it a living being"[10]. For our purposes, we will simply view human physiology as man's description of his own bodily processes. Since the computer is beginning to affect almost every aspect of our lives, could the computer's capabilities provide us with useful information concerning the body's physiological mechanisms?

We already have CT[1] scanners which provide us with detailed images of pathophysiological effects. Clearly this is extremely useful, but from an anthropocentric viewpoint is simply augmentation of the visual input to the physician's decision making procedures. Automated ECG analysis is now commonplace, but human pattern recognition of ECG forms still remains superior to that of the computer and for now this problem of monitoring and interpreting electrical propagation within the heart remains within the domain of signal analysis.

Let us go one step further and ask ourselves whether we could somehow represent in the computer the physician's *mental model* of physiology. If such a computer program were created, we could use it to "intelligently" augment the physician's power of reason. Nils Nilsson describes such programs as follows:

> "Many human mental activities such as writing computer programs, doing mathematics, engaging in commonsense reasoning, understanding language, and even driving an automobile are said to demand 'intelligence'. Over the past few decades, several computer systems have been built that can perform tasks such as these.
> ... We might say that such systems possess some degree of *artificial intelligence*."[17]

In essence we have stepped into the research area of Artificial Intelligence in Medicine.

---

1. Computerized *t*omography. The conjunction of an x-ray machine to a computer.

If a computer program can model what goes on in the mind of a physician when he or she thinks in physiological terms, what properties would it possess? Just as any other program, it takes in data as input, performs a computation and produces some form of output. In terms of the program NEPHROS presented in this thesis, we will call the input data $\mathcal{M}$, the set of manifestations[1] of some disorder in the particular patient $\mathcal{P}$ currently under consideration. As an example, $\mathcal{M}$ could be the set {decreased cardiac output, increased sodium retention}.

The computation performed by the program will be discussed in detail later. For now it will simply be described as propagation of the given information, $\mathcal{M}$, within a network of causal relationships. These relationships are "facts" such as "an increased level of antidiuretic hormone in the collecting tubule of the nephron causes an increased reabsorption of fluid at that site." The output produced by the program is $\mathcal{N}$, the set of new information derived from propagating the input data. $\mathcal{N}$ could be, for example, {decreased blood pressure, increased aldosterone secretion, increased interstitial fluid}.

The causal network has been designed so that most of the new information derived by NEPHROS is clinically relevant. Therefore, is the job of the computer program finished when it prints out a list of the members of set $\mathcal{N}$? Surely we could not expect a physician to accept the results from a computer program on faith alone. An example will help illustrate this point.

1.2 Guyton's Model Of Circulatory Physiology

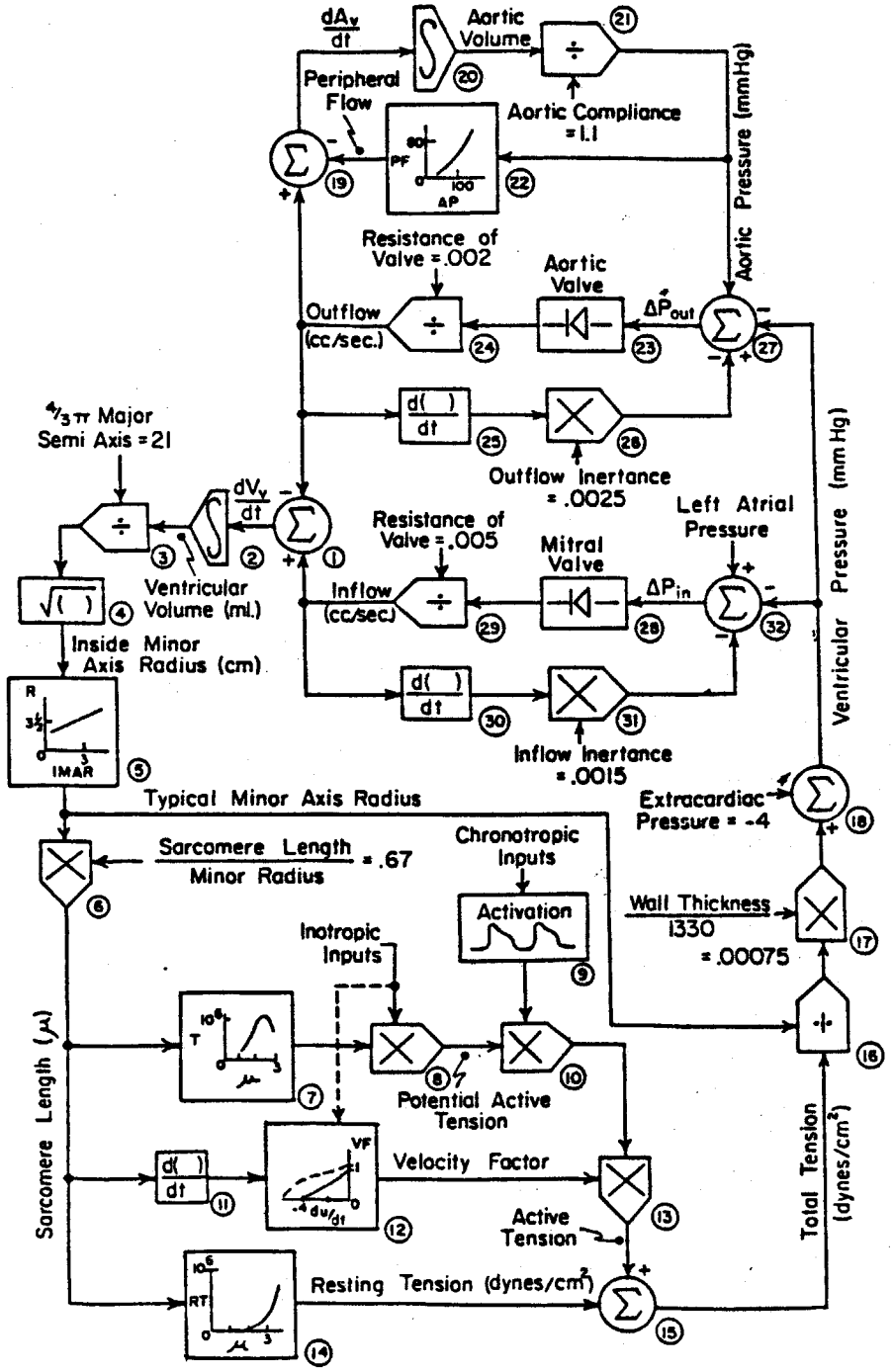
To make tangible his belief that "the human being is actually an automaton",[2] Guyton has created a computer program to model total circulatory function[11]. The final simulation network he produced is truly ingenious. A part of it is shown in figure 1. He used standard mathematical operations, integration and graphical techniques to represent the relationships between physiological variables. For example, at the top of the diagram we can see how the equation,

$$\text{Compliance} = \Delta\text{Volume} \div \Delta\text{Pressure}$$

---

1. Signs and symptoms.
2. "The very fact that we are alive is almost beyond our own control, for hunger makes us seek food and fear makes us seek refuge. Sensations of cold make us provide warmth, and other forces cause us to seek fellowship and to reproduce. Thus the human being is actually an automaton ..."[10]

Fig. 1. Guyton's Model of Left Ventricular Function



A systems analysis in block-diagram form, depicting function of the left ventricle and of the immediately adjacent circulation.

is represented. To simulate the passage of time, the program has an internal "clock" which is incremented at small, well-defined time units. At each tick of the clock, the physiological variables are adjusted using the mathematical relationships represented in the simulation network.[1]

Let us assume for a moment that we have constructed a very large simulation network of the kind shown in figure 1 which accurately models the whole human body. Such a program would be invaluable. We could watch the physiological variables change over time while we speed up slow processes or slow down fast processes. We could extrapolate into the future by allowing our computer program to simulate in seconds what would be days or months in real life. We could even give negative increments to the "clock" to run the program backwards in time. In this manner we could determine the values of the physiological variables at times in the past.

But if such a program existed, how would we justify our results? With hundreds or even thousands of interdependent mathematical equations determining the manner in which the physiological variables change, this would be a formidable task. Shortliffe addressed the importance of advice explanation when describing MYCIN, a computer program which provides infectious disease consultation:

> "... a physician will be more willing to accept a program's advice if he is able to understand the decision steps that the system has taken. This gives him a basis on which to reject the system's advice if he finds that the program is not able to justify its decisions sufficiently. It thereby helps the program conform to the physician's requirement that a consultation system be a tool and not a dogmatic replacement for the doctor's own decisions."[20]

Therefore, it is crucial for a computer program modelling any aspect of medicine to have the ability to explain how it has derived results. For this reason, the explanation facility of NEPHROS has been intimately interwoven with the program's computational apparatus.

---

1. Guyton uses a clock-driven simulation. As will be discussed later, Johan de Kleer's conjunction of envisionment and propagation of constraints has been implemented in NEPHROS. The combination of these two techniques produces an abstract event-driven simulation.

1.3 The NEPHROS Explanation Facility

The main computational structure in NEPHROS is a strict hierarchy of independent elements. The tasks performed by each of these devices is meant to model the workings of some physiological entity, such as the kidney. There is no need for these entities to be anatomically concrete; the program has computational devices representing abstract concepts such as oncotic pressure. At all levels of the hierarchy but the top, each computational device has been defined to be a subcomponent of another at a higher level. For example, the "kidney" is composed of devices corresponding to the juxtaglomerular apparatus, glomerulus, proximal tubule, distal tubule and collecting tubule.

It would be completely unfair and unnecessary to keep this information hidden from someone interacting with the program. With this belief in mind, NEPHROS was designed to give the user access to all components of the hierarchy. This permits one to examine the input/output behaviour of the "kidney" and then remove its cover and examine the input/output behaviour of the "collecting tubule." By passing up or down the computational hierarchy, values of physiological variables can be given in the context in which they truly occur. All this without a biopsy!

1.4 Physiology and Reductionism

Physiology has been neatly classified into systems such as respiratory, cardiovascular, neurological, musculoskeletal, hematological and so forth. This modularity is carried further when we describe physiological systems in finer detail. For example, nervous system function can be segmented into those processes occurring at either the spinal cord level, the lower brain level[1] or the higher brain (cortical) level. What we end up with is a conceptual hierarchy. This description methodology is termed *reductionism*,[2] the assumption that something can be understood if the nature of its subcomponents is known. Medical reductionism is often naturally led by anatomical localization of function. As an example, the reticular substance of the medulla

---

1. Medulla, pons, mesencephalon, hypothalamus, thalamus, cerebellum and basal ganglia[10].
2. Hofstadter provides an entertaining treatise on the reductionism versus holism argument in [12].

and pons of the lower brain level are the main centers for subconscious control of arterial blood pressure and respiration.

It must be said that this functional hierarchy does not occur within physiological contexts alone and is found throughout almost all medical literature. This technique has proven to be very useful in modelling real world medicine. However, can we conclude that reductionism is powerful enough to capture all aspects of the body's mechanisms? Perhaps not. In a reductionistic attempt to model the brain using *perceptrons*, Marvin Minsky and Seymour Papert noted some restrictions:

> "... the image is that of a network of relatively simple elements, randomly connected to one another, with provision for making adjustments of the ease with which signals can go across the connections. ... A perceptron whose Φ's are properly designed for a discrimination known to be of suitably low order will have a good chance to improve its performance adaptively. Our purpose is to explain why there is little chance of much good coming from giving a high-order problem to a quasi-universal perceptron whose partial functions have not been chosen with any particular task in mind[15]".

The program presented here does use simple computational devices, but they certainly will not be burdened with high-order problems. To represent the renal physiology[1] underlying three syndromes, a hierarchy of simple computational devices shall be introduced. I conjecture that for other physiological systems of *similar complexity*, we should not run into insurmountable difficulties using that methodology.

1.5 Why Renal Physiology?

The original goal of this project was to develop a computer model for some aspect of human physiology. The resultant program was to provide useful justifications of new physiological information produced. Renal physiology was chosen primarily because it possesses a complex and fascinating repertoire of causal interactions which indeed provided a challenge for the modelling and explanation problem. It should be said, however, that the functions of the kidney are many. NEPHROS was constructed to model the renal physiology underlying three syndromes

---

1. The subset of human physiological mechanisms which pertain to the kidney.

which required representation of the causal dependencies underlying sodium and water balance alone.

Renal physiology, as presented here, is essentially a hormonal system and therefore is dependent upon the circulation of blood for its function. This provided the impetus for building a representation for the full circulatory loop of the body.[1] In physiological domains other than that pertaining to the kidney, the blood is again an important route for the passage of information. It is hoped that future attempts at modelling physiological systems will be facilitated by the attempt here at capturing this anatomical feature.

Patil and Szolovits of M.I.T. have constructed ABEL, a computer program which provides expert consultation for the subset of renal physiology dealing with acid-base and electrolyte disturbances[18]. ABEL has the ability to express causal relationships between physiological entities, but these are high-level and associational in nature. For instance, a typical relationship found at the pathophysiological level is "low-pH" causes "potassium-shift-out-of-cells". Such a statement leaves unanswered the question "By what mechanism does this occur?". In other words, there is an even deeper level of causal description which corresponds to details of physiological mechanism. One can envisage the passage of information between ABEL and a computer program similar to NEPHROS in which acid-base metabolism has been represented. Thus a computer model of physiology can be useful to both man *and* machine.

1.6 Significance

How is the construction of a computer program like NEPHROS useful to the research area of artificial intelligence in medicine?

1. To help us formalize our ideas about the causal relationships in physiological mechanisms.

---

1. Blood is carried from the heart and passes through the blood vessels, to the tissues and back to the heart.

2. Since physiological function is often physically localized, modelling bodily processes prods us to represent *anatomical* relationships in a computer program.

3. Such a computer program can be clinically useful. If the manifestations of a particular patient have been supplied to it, hypotheses can be tested, known information verified and perhaps even prognostic information can be provided.

4. As mentioned earlier, the program could interact with programs like ABEL which model medicine at a higher level of abstraction.

What purpose does NEPHROS serve as an artificial intelligence construct?

1. As an application of the techniques of qualitative reasoning, constraint propagation and envisionment.

2. The introduction of a technique to be described later, the use of *explanation boxes*. This allows:

   (a) *Justification* of information derived from propagation within a hierarchical constraint network.

   (b) Specification of *purpose*.

   (c) Dynamic construction of constraint networks at run time.

2. Interacting With NEPHROS

So that the reader will have a more concrete understanding of the task performed by NEPHROS, an interaction with the program is presented below. The rest of the thesis describes the internal workings of the program to explain how it possesses the features demonstrated in this chapter.

Figure 2 illustrates the corresponding propagation of information within NEPHROS and may be used as a guide for this chapter. The numbers 1 and 2 label the first and second passes through the body's circulatory loop, respectively. In the dialogue below, user input is prompted by "⇒". Important notes pertaining to the interaction are given in **bold face.**

2.1 Input Phase

During the first stage of input, the program asks if the values of any physiological variables are known. As we shall see later, all variables occur within a specific physiological context. For example, the variable sodium retention is associated with the kidney.

Welcome To NEPHROS! If you wish to enter any information, please type the corresponding physiological context. Type END when you wish the input phase to halt.

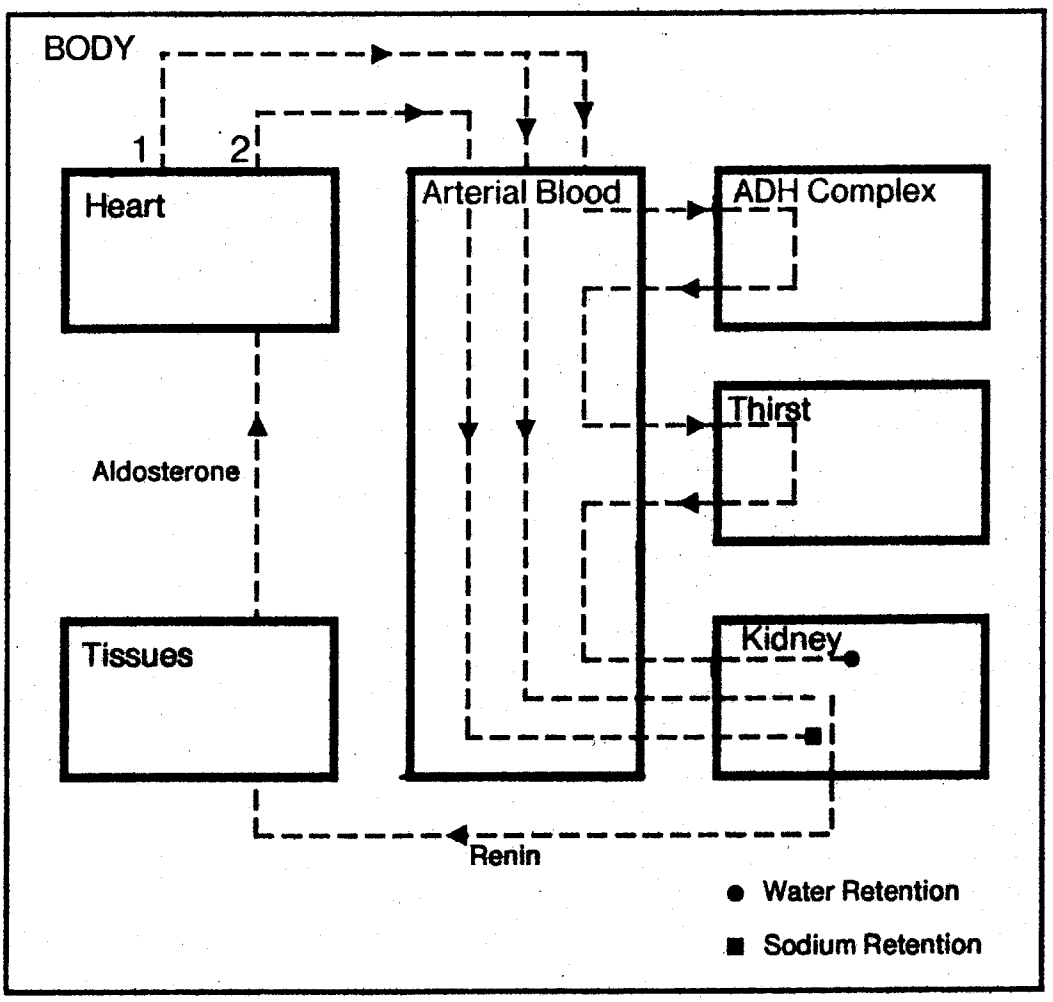
⇒ ARTERIAL BLOOD

Type in an assignment to a physiological variable pertaining to the ARTERIAL BLOOD. Input should be of the form MEASURE VARIABLE.

⇒ NORMAL TOTAL PERIPHERAL RESISTANCE

NEPHROS keeps asking for value assignments to physiological variables until END is typed in.

Fig. 2. Information Propagation in Heart Failure



If you wish to enter any information, please type the corresponding physiological context. Type END when you wish the input phase to halt.

⇒ END

The second stage of input involves typing in hypotheses concerning the patient's condition. Corresponding to each hypothesis, there is a set of physiological assignments which the program will assume.

You may incorporate an hypothesis concerning the patient's condition into the physiological constraint network.

⇒ ?

? is typed in when the user desires a list of available choices. We will have the program determine the consequences of heart failure.

HEART FAILURE
SIADH
NEPHROTIC SYNDROME
NO HYPOTHESIS

⇒ HEART FAILURE

Since the hypothesis is HEART FAILURE, a DECREASED CARDIAC OUTPUT shall be assumed. Should this assumption be made?

⇒ YES

Since heart failure was chosen above, NEPHROS has replaced the normal HEART by HEART IN FAILURE in its representation of renal physiology. This is an example of how the program adjusts its model of physiology during the input phase. The activation of the HEART IN FAILURE component caused a DECREASED CARDIAC OUTPUT to be assumed.

2.2 Propagation Phase

The above input information is propagated in a network of physiological relationships. This propagation will be described in detail in chapter 6.

In order to determine ADH SECRETION of the ADH COMPLEX, BLOOD OSMOLARITY of the ADH COMPLEX should be known. Please enter a value for that physiological variable.

⇒ ?

INCREASED
DECREASED
NORMAL
UNKNOWN

⇒ NORMAL

The program has determined that arterial blood pressure is decreased. The ADH Complex needed this fact plus the information concerning blood osmolarity to determine ADH secretion. Since knowledge of ADH secretion is required to permit further information propagation, the program has been constructed to ask for osmolarity in this circumstance. For a similar reason, the program asks for osmolarity of the blood flowing through the thirst complex:

Should a value of NORMAL also be assigned to BLOOD OSMOLARITY of the THIRST COMPLEX? A value for that physiological variable is needed in order to determine WATER INTAKE of the THIRST COMPLEX.

⇒ YES

NEPHROS asks for data at other critical points during the information propagation. Here is an example of how the program changes the building blocks of its model of physiology during the propagation phase:

INTERSTITIAL FLUID of the TISSUES is INCREASED. Should an excessive amount of fluid in the tissues (edema) be assumed?

⇒ YES

The normal TISSUES model has been replaced by EDEMATOUS TISSUES in the NEPHROS physiological network.

2.3 Explanation Phase

After all information propagation has ceased, NEPHROS allows the user to examine new knowledge derived within its model of physiology. This model, a network of physiological relationships, is constructed much like the human body. To illustrate how we can access the new data, we shall ask the program for the value of kidney water excretion.

You are now in the NEPHROS explanation phase. Please enter a physiological context to be examined.

⇒ ?

BODY
HEART
TISSUES
ARTERIAL BLOOD
ADH COMPLEX
THIRST
KIDNEY
JUXTAGLOMERULAR APPARATUS
...

⇒ BODY

Which physiological variable pertaining to the BODY do you wish to examine?

⇒ ?

WATER INTAKE

KIDNEY SODIUM EXCRETION
KIDNEY WATER EXCRETION

⇒ KIDNEY WATER EXCRETION

KIDNEY WATER EXCRETION of the BODY is DECREASED. Do you wish further explanation?

We are now in a position to ask the program to justify why kidney water excretion is decreased. Let us get a list of the commands we can use to travel within the program's physiological model. The commands will be explained by example.

⇒ ?

RESTART
DOWNWARDS
UPWARDS
SAME LEVEL
QUIT

We are currently examining within the BODY causal context. Figure 2 shows that this environment encompasses all others. Let us attempt to go to a more general context which includes the BODY.

⇒ UPWARDS

You are already examining at the highest level of description, the BODY physiological context.

From the BODY context we can only go down to finer levels of description. Since NEPHROS associates water excretion with the kidney, a step downwards puts us in the KIDNEY causal context.

⇒ DOWNWARDS

KIDNEY WATER EXCRETION of the BODY is DECREASED since WATER EXCRETION of the KIDNEY is DECREASED. Do you wish further explanation?

We are now at the level of description corresponding to the information propagation

shown in figure 2. Let's justify kidney water excretion by tracing back some of the propagation pathways the program previously created.

⇒ SAME LEVEL

WATER EXCRETION of the KIDNEY is DECREASED since
1. BLOOD FLOW of the KIDNEY is DECREASED.
2. ADH LEVEL of the KIDNEY is INCREASED.

⇒ SAME LEVEL

Which causal influence?

Two information pathways converged to determine water excretion. In this case SAME LEVEL is ambiguous and we must specify which of the two causal influences to consider.

⇒ 2

ADH LEVEL of the KIDNEY is INCREASED since ADH LEVEL of the ARTERIAL BLOOD is INCREASED.

⇒ SAME LEVEL

Explanation is now within the ARTERIAL BLOOD causal context.

ADH LEVEL of the ARTERIAL BLOOD is INCREASED since ADH SECRETION of the ADH COMPLEX is INCREASED.

Explanation is now within the ADH COMPLEX causal context. If desired, we could have examined how the decreased arterial blood pressure associated with heart failure influenced ADH secretion. Instead, we'll examine the effect of ADH on the collecting tubule of the kidney as an illustration of how to restart the explanation phase. Notice below that NEPHROS sometimes includes purpose information in its explanations.

⇒ RESTART

Please enter a physiological context to be examined.

⇒ KIDNEY

Which physiological variable pertaining to the KIDNEY do you wish to examine?

⇒ WATER EXCRETION

WATER EXCRETION of the KIDNEY is DECREASED since

1. BLOOD FLOW of the KIDNEY is DECREASED.
2. ADH LEVEL of the KIDNEY is INCREASED.

⇒ DOWNWARDS

WATER EXCRETION of the KIDNEY is DECREASED since WATER REABSORPTION of the COLLECTING TUBULE is INCREASED.

⇒ SAME LEVEL

WATER REABSORPTION of the COLLECTING TUBULE is INCREASED since ADH LEVEL of the COLLECTING TUBULE is INCREASED. The purpose of this is to restore to normality the effective arterial blood volume which is currently DECREASED.

⇒ QUIT

Bye.

3. Relevant AI Techniques

In later chapters, the workings of NEPHROS are described in full detail. To enable this, the relevant artificial intelligence tools used by the program are introduced here.

3.1 Envisionment

In order to represent the thinking processes of physicians, we will have to make rational approximations to what goes on in their minds. In particular, what mental model does the physician use when thinking in physiological terms? For example, in heart failure there is a sequence of events beginning with decreased cardiac output and culminating in both increased sodium retention and increased fluid retention. How is this played out in the physician's mind?

This concept should not seem strange to us since we perform similar mental acrobatics every day. For sake of illustration, let us make an attempt to describe the sequence of events starting from the production of sound from a lecturer's vocal cords and finishing with vibration of a student's eardrum. We will assume that the lecturer has been supplied with a public address system. To keep things simple, we will not worry about sound resonance within the body nor amplification of voltage signals from the microphone.

Vibration of the lecturer's vocal cords produces sound waves. They pass through the air and enter the microphone. The microphone converts the sound waves into electrical impulses. These voltage signals pass into the speaker which then casts sound waves into the air of the auditorium. These waves reach the student's eardrum, resulting in its vibration.

Johan de Kleer introduced the term *envisionment* for this qualitative simulation of a series of events[5][6]. There are some important points we should note about this kind of simulation:

1. As is the case for computer simulation, there may be no direct correlation between the time elapsed during the envisionment and the amount of time it takes for the process to occur in reality. However, there is a direct mapping between events described in the envisionment and those which actually occur.

2. This type of simulation is often imprecise. The exact time elapsing between events is frequently not specifically mentioned. In addition, causal mechanisms are often not completely described. For example, from the above description we do not know *how* the sound waves are converted by the microphone into electrical impulses.

3. Purpose is not explicitly mentioned. For the envisionment above, it is useful information that the microphone was specifically designed to convert sound waves into electrical impulses.

4. Envisionment *does* allow us explain how certain results were derived. A question such as "How was the student's eardrum set in motion?" could be easily answered by tracing backwards the mental pathways which were previously created.

With a little massaging, envisionment was instrumental in providing a basis for the computer program described here. The assumption was made that doctors often dynamically create these mental descriptions to rationalize physiological behaviour. In a later chapter we shall see how NEPHROS performs its own envisionment to parallel what goes on in the mind of a physician reasoning in physiological terms.

3.2 Qualitative Reasoning

In the envisionment given in the previous section, we did not once mention quantitative values. Statements such as "the passage of sound to the microphone caused an electrical impulse of 5 millivolts to be generated" were not used. In order to model the envisionment process in a computer program and to be able to explicitly refer to values of *qualitative* variables, we will have to use some formal qualitative algebra.

The algebraic system which we will use is called *Incremental Qualitative* (IQ) algebra, introduced by de Kleer[4]. Here we will label the qualitative values *normal, increased* with respect to normal, *decreased* with respect to normal and *unknown*.[1] To be precise, we will have to specify what is meant by *normal*. From here on normality will be with respect to the *particular patient* whose physiology is being modelled. When "increased aldosterone secretion" is given to

NEPHROS, that could mean 180 μg/24 h in one patient and 250 μg/24 h in another, depending upon which patient is being modelled.

3.3 Propagation of Constraints

What is the best method for determining the set $\mathcal{N}$ of new information given the set $\mathcal{M}$ of manifestations and a network of causal dependencies? Similar artificial intelligence problems have been efficiently mastered using the method of *propagation of constraints*. Sussman and Stallman used constraint propagation as the basis of a system for computer-aided circuit analysis [22] [23]. Later research by Sussman and Steele focused on the development of a general language to represent constraints [24] [21]. Johan de Kleer united qualitative analysis and propagation of constraints in the development of a theory of human-like reasoning within the domain of electronic circuit analysis and recognition [6].

Electronic circuit theory provided a natural and fruitful testing ground for the technique of propagation of constraints. Waltz applied this range constriction procedure to assign Huffman-Clowes junction and line labels to given scenes and therefore constraint propagation proved to be applicable to other areas of interest[28]. As we shall see, human physiology contains many causal and associational relationships that may be represented as constraints.

The subdomain of circulatory physiology provides many examples of constraints. This is by no means an accident since there is a direct parallelism between fluid dynamics and electronic circuit theory. The fact that blood pressure equals the product of cardiac output and total peripheral resistance is nothing other than a restatement of Ohm's law. The fact that the amount of blood entering the bifurcation of the brachial artery equals the sum of the amounts of blood entering the two output vessels (radial and ulnar arteries) is simply a variation of the Kirchoff Current Law constraint. If one considers the relative constriction of the radial and ulnar arteries, the bifurcation can be thought of as a current divider – another familiar constraint.

---

1. de Kleer actually uses *not changing, increasing, decreasing and unknown*. The use of these terms here may confuse the difference between the value of a variable and the sign of the derivative of the variable. In addition, I believe that the labelling given above is more natural for most medical explanations.

Fortunately, renal physiology provides us with many relationships which may be represented in terms of constraints. In a discussion with Dr. Schwartz[1] of Tufts-New England Medical Center, he described step-by-step the physiology behind SIADH (*s*yndrome of *i*nappropriate secretion of *anti*diuretic *h*ormone), one of the syndromes modelled by NEPHROS [1]. An example of the qualitative constraint reasoning used by Dr. Schwartz is given below:

> Sodium excretion is influenced by glomerular filtration rate, aldosterone and the mysterious third factor [natriuretic hormone]. An increased GFR means a greater amount of sodium per unit time entering the tubule and thus increased sodium excretion. Aldosterone decreases sodium excretion by increasing distal nephron absorption of sodium. The third factor is believed to be a hormone which increases sodium excretion by blocking reabsorption of sodium in the proximal part of the nephron.

We can think of GFR, aldosterone and natriuretic hormone as "forces", each attempting to affect sodium excretion in their own particular manner.

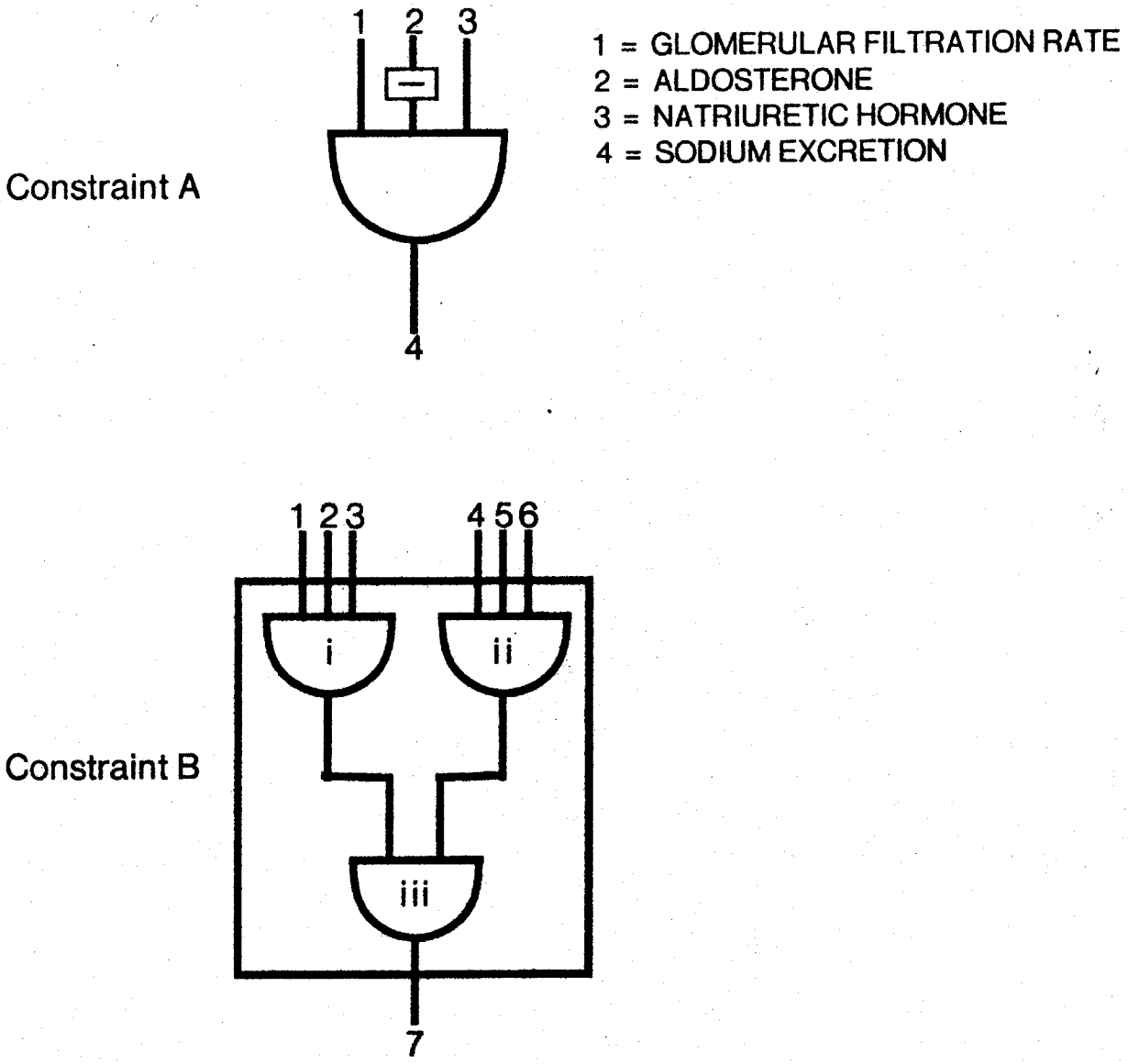
The above relationship can be represented as a *constraint* using IQ algebra. In a qualitative sense, glomerular filtration rate, aldosterone and natriuretic hormone *add* together to determine sodium excretion. The adder is illustrated in Figure 3 and labelled "Constraint A". The first input to the adder, which we will call pin 1, corresponds to glomerular filtration rate (GFR). Pins 2 and 3 correspond to the level of aldosterone in the blood and the influence contributed by natriuretic hormone, respectively. Pin 4 is the final "sum" of this adder, corresponding to sodium excretion. Since aldosterone has a negative influence on sodium excretion, an "IQ-inverter" is placed between pin 2 and the adder.

Let us put constraint A into action. Assume glomerular filtration rate is *normal*, aldosterone level is *decreased* and natriuretic "hormone" is *increased*. Since GFR is normal, it plays a passive role and it will be the other two pins which will assign a qualitative value to sodium excretion. Before entering the adder, the value of the aldosterone level is inverted to *increased*. If the IQ-adder has been appropriately constructed, the input values of *increased* contributed by aldosterone and *increased* contributed by natriuretic hormone "sum" to *increased*, the new value for sodium excretion.

---

1. Dr. W.B. Schwartz, University Professor of Medicine, Tufts University.

Fig. 3. Examples Of Constraints



What value would be assigned to sodium excretion if the values of GFR and aldosterone are left as they are, but the initial value of natriuretic hormone is set to *decreased*, rather than *increased*? Since we have a value of *increased* from pin 2 after the inversion and *decreased* from pin 3, what is the sum? In this situation, sodium excretion is given the value *unknown*. Using similar deduction, we can fully delineate the results of adding any two IQ values.

If we keep pin 1 set to *normal*, and thus converting the three-input adder into a two-input adder, what would be the results of all combinations of pin values? This is answered in the table below. These definition tables get rather large; for a three-input IQ adder, there would be 64 rows. It is therefore understandable why GFR was artificially set to *normal* for this illustration. However, it should be mentioned that although the tables are massive, the rules necessary and sufficient to compute IQ values are rather small in number.

Fig. 4. Two-Operand IQ Addition

	PIN I	PIN II	SUM
1	INCREASED	INCREASED	INCREASED
2	INCREASED	NORMAL	INCREASED
3	INCREASED	DECREASED	UNKNOWN
4	INCREASED	UNKNOWN	UNKNOWN
5	NORMAL	INCREASED	INCREASED
6	NORMAL	NORMAL	NORMAL
7	NORMAL	DECREASED	DECREASED
8	NORMAL	UNKNOWN	UNKNOWN
9	DECREASED	INCREASED	UNKNOWN
10	DECREASED	NORMAL	DECREASED
11	DECREASED	DECREASED	DECREASED
12	DECREASED	UNKNOWN	UNKNOWN
13	UNKNOWN	INCREASED	UNKNOWN
14	UNKNOWN	NORMAL	UNKNOWN
15	UNKNOWN	DECREASED	UNKNOWN
16	UNKNOWN	UNKNOWN	UNKNOWN

Does the table capture all the information about two-input IQ addition? Let us glance at row 4. It tells us that if pin I of a two-input IQ-adder is *increased* and pin II is *unknown*, then the "sum" is *unknown*. But what if we know that the sum is *decreased*? In order for pin I to have the value *increased* and the sum to be *decreased*, the value of pin II would have to be *decreased*. In other words, we lose information by making a constraint unidirectional. To make our definition of the two-input IQ adder complete, there should be a table describing how the value of pin II can be derived if the value of pin I and the sum are known. Similarly, there should be a table describing how the value of pin I can be derived if the value of pin II and the sum are known. A two-input IQ-adder can thought of as a black box with three pins; setting two of the pins will give us a value, perhaps *unknown*, for the third pin.

We are not restricted to constructing IQ-adders or IQ-inverters. To represent the fact that blood pressure is the product of cardiac output and total peripheral resistance, an IQ-multiplier should be used.[1] Any qualitative relationship could be defined and the corresponding constraint created. However, it should be noted that caution must be exercised when multidirectional constraints are constructed. If we assume that "heart failure causes decreased cardiac output", we can't immediately assume that "a decreased cardiac output means there is heart failure".

Constraint B of Figure 3 shows how three adders can be combined to create a larger constraint. If the values for pins 1 through 6 are known, one can envisage the sums of adders i and ii *propagating* to adder iii and generating a result at pin 7. This process is called *propagation of constraints*. The skeleton of NEPHROS consists of constraints representing causal relationships upon which the elements of the set $\mathcal{M}$ are propagated. The previous table demonstrated that many relationships can be compacted into a single constraint. It is quite easy to imagine that large constraint networks can contain a great deal of information.

As mentioned earlier, de Kleer developed a theory of human-like reasoning within the domain of electronic circuit analysis and recognition[6]. To carry out this task, he united the techniques of qualitative analysis and propagation of constraints. The goal here was to determine if this combination could naturally be applied to renal physiology, with particular concern given to

---

1. IQ-multipliers and IQ-dividers have identical definition tables as IQ-adders and IQ-subtractors, respectively.

the quality of explanations produced by the program. Doctors most likely use pure qualitative reasoning for a large part of their decision-making and thus the use of such reasoning in computer-derived explanation should pose no major difficulty. However, constraint propagation provides an efficient and explicit representation for relationships, rather than pedagogical elegance. The major thrust of the research presented here is to justify the viability of the combination of constraint theory and qualitative analysis for explanation of physiological concepts.

3.4 Grey Boxes

Consider again constraint B of Figure 3. Note the square surrounding the constraint. Let us imagine that all the area inside the square has been painted black. If the values for pins 1 through 6 are known, we can still derive a value for pin 7. But what information has been lost? Since neither the adders nor their interconnections are visible to us, an explanation for *how* a value for pin 7 is derived cannot be given. If the goal is to derive detailed explanations, these *black boxes* will not suffice.

It should be noted, however, that there are circumstances in which it is convenient to ignore subcomponents and their interconnections. The user of a complex integrated circuit is usually not interested in a detailed description of the network of logic gates which make up the device. Also, the black box approach often facilitates creation. From the manufacturer's viewpoint, once an efficient design for an integrated circuit has been made, the circuit can be mass-produced in one piece without having to fit together subcomponents which were created individually.

It should be emphasized that black boxes are not totally inadequate for explanatory purposes, they are inadequate for the generation of *detailed* explanations. We certainly can rationalize why a person with excessive urine output is thirsty without referring to the effect of increased osmotic pressure on the lateral preoptic area of the hypothalamus!

Consider again the issue of modelling renal physiology. An important point which surfaced in the discussion with Dr. Schwartz was that he changed the level of detail in his explanations depending upon the demands of the concept being discussed. He mentioned that the intricate details of the nephron were not referred to in most pathophysiological situations, but were

important in some instances. In certain situations it is unimportant to know that natriuretic hormone acts on the proximal convoluted tubule of the nephron. In the same circumstances a computer program modelling the reasoning process should also "know" this information is unimportant. Thus the ability for the computer program being described here to *vertically* change the level of detail in the explanation phase will be crucial. In other words, any constraints above the lowest level of detail should be represented as *grey boxes*[23].

3.5 Horizontal Abstraction

In the previous section, the importance of a system to generate descriptions at various levels of detail was emphasized. This will be termed *vertical abstraction*. In this section we shall see the importance of *horizontal abstraction*. The distinction is made between three types of horizontal shift of abstraction:

1. *Slices*. Descriptions of some object which may coexist.

2. Mutually exclusive descriptions of some object.

3. Changing the focus of concentration from some object to another at the same level of detail. We will term this *shift of focus*.

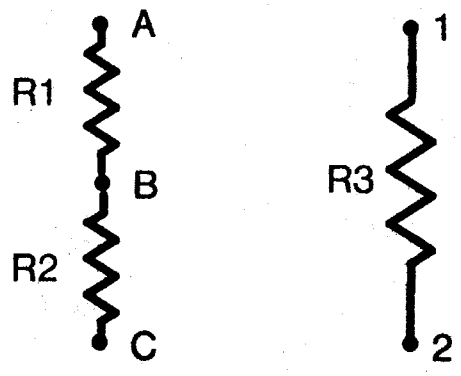
3.5.1 Slices

Consider the resistor configurations below. Assume we know the values of the resistances R1 and R2, and the voltages V_A and V_C with respect to ground. What is the current through R1 and R2 ? Since we don't know V_B, local application of Ohm's Law[1] to R1 or R2 will not give us the answer. However, since R1 and R2 are in series, we know that the resistance across terminals A and C is the sum of R1 and R2. Let this sum be R3. If we let V_1 assume the value of V_A and V_2 assume the value of V_C, we can directly apply Ohm's law to R3 and immediately derive the value

1. The voltage drop across a resistor equals the product of current through the resistor and the resistance.

of the current through R3. Since R3 was just another way of looking at the R1-R2 combination, the current through R1 and R2 is now known.

Fig. 5. Slices



By introducing a different representation for the same object, we derived information that ordinarily we would not have been able to obtain. Sussman refers to R3 and the R1-R2 combination as *slices.* In his own words:

> "... slices are redundant descriptions – the same truths from a variety of viewpoints. The way that the slices do their job, however, is by providing redundant paths for information to travel in the process of analysis."[25]

The preliminary version of the NEPHROS constraint network presented later does not contain slices. Redundant pathways for information propagation are currently being designed. The purpose of introducing slices here is to allow them to be contrasted with mutually exclusive descriptions, which are introduced below.

3.5.2 Mutually Exclusive Descriptions

Slices are essentially coexistent representations of the same object. Medicine contains many such relationships. For instance, albumin may be viewed as both a major blood osmole and as a Ca^{2+} carrier. Medicine also provides us with a wealth of *conflicting* viewpoints. A disease is by definition a disorder of bodily function and hence provides viewpoints apart from the norm.

When the physician finds a massive amount of protein in the urine, in his or her own mind's eye a normal glomerulus is replaced by one in which permeability of the filtration membrane has increased. After this mental building block has been set in place, the new information is propagated. If no strong contradiction is found, an hypothesis is generated and the leaky membrane hides behind the guise of *nephrotic syndrome*. It is this process that the program presented here attempts to model.

3.5.3 Shift Of Focus

When the propagation of values within constraint B of figure 3 was discussed, we implicitly concentrated first on adder i, then adder ii and then adder iii. For a computer we must be explicit and formalize the change of concentration from one object to another.

It should be noted that all three adders of constraint B are at the same level of description. At a finer level of detail there are the internal workings of the adders. At a less precise level of description, we have the "black box" described earlier. A horizontal shift within a particular plane of description shall be termed *shift of focus*.

4. Explanation Boxes

4.1 Causality and Computation

The previous chapter demonstrated the plausibility of representing renal physiological relationships in a constraint network. Since de Kleer had considerable success with uniting qualitative reasoning and propagation of constraints, we should not expect to run into extreme difficulties when modelling the envisionment process with such a network. However, will explanation of newly derived information be a hard task? We are modelling causality and therefore justifications for the program's behaviour should be given in those terms. To a physician, "pin 1 of constraint 23 which is contained in constraint 9 has the value *increased*" provides no useful information whatsoever.

Let us reexamine the envisionment presented earlier which described the sequence of events beginning with a lecturer's sound production and ending with vibration of a student's eardrum. First let us put our minds within the context of the vocal cords. Sound generation can be thought of as an *output* of the vocal cord complex. Now consider the microphone. The sound *information* from the vocal cords is relayed by the air to the microphone. In the context of the microphone, the sound is *input* and the voltage signals produced are *output*. The information contained in the electrical impulses is then carried to the speaker. In the context of the speaker, these signals are taken as input and converted to sound waves. The sound waves are carried by the air to a student's ear and can be looked upon as input to the eardrum.

4.2 Purpose

In the above discussion, a deliberate attempt was made to present each "context" as a computational device. These contexts were described as having inputs, performing a computation and having outputs. The assumption is made here that we can capture the computational component using a constraint representation. But is there other information lying in these contexts that would be of use to us?

The importance of *purpose* information has been stressed. In the scenario above, vocal cords were used to create sound waves. The microphone was purposely constructed to convert sound waves into electrical impulses. The speaker and eardrum too have purposes. Man-made objects are usually created with some particular task in mind. Without getting into deep philosophical or theological issues, let us assume that the physiological objects to be described later are constructed and connected with some purpose "in mind". They certainly function in that manner.[1]

To mesh neatly with causal computations, we have to define *purpose* in general terms. For example, although we know an increased level of ADH causes increased water reabsorption in the collecting tubule for the purpose of restoring blood volume, we also know that a decreased level of ADH causes decreased water reabsorption in the collecting tubule for the purpose of lowering blood volume. Therefore, there should be *dynamic purpose computation*.

4.3 Activation Conditions

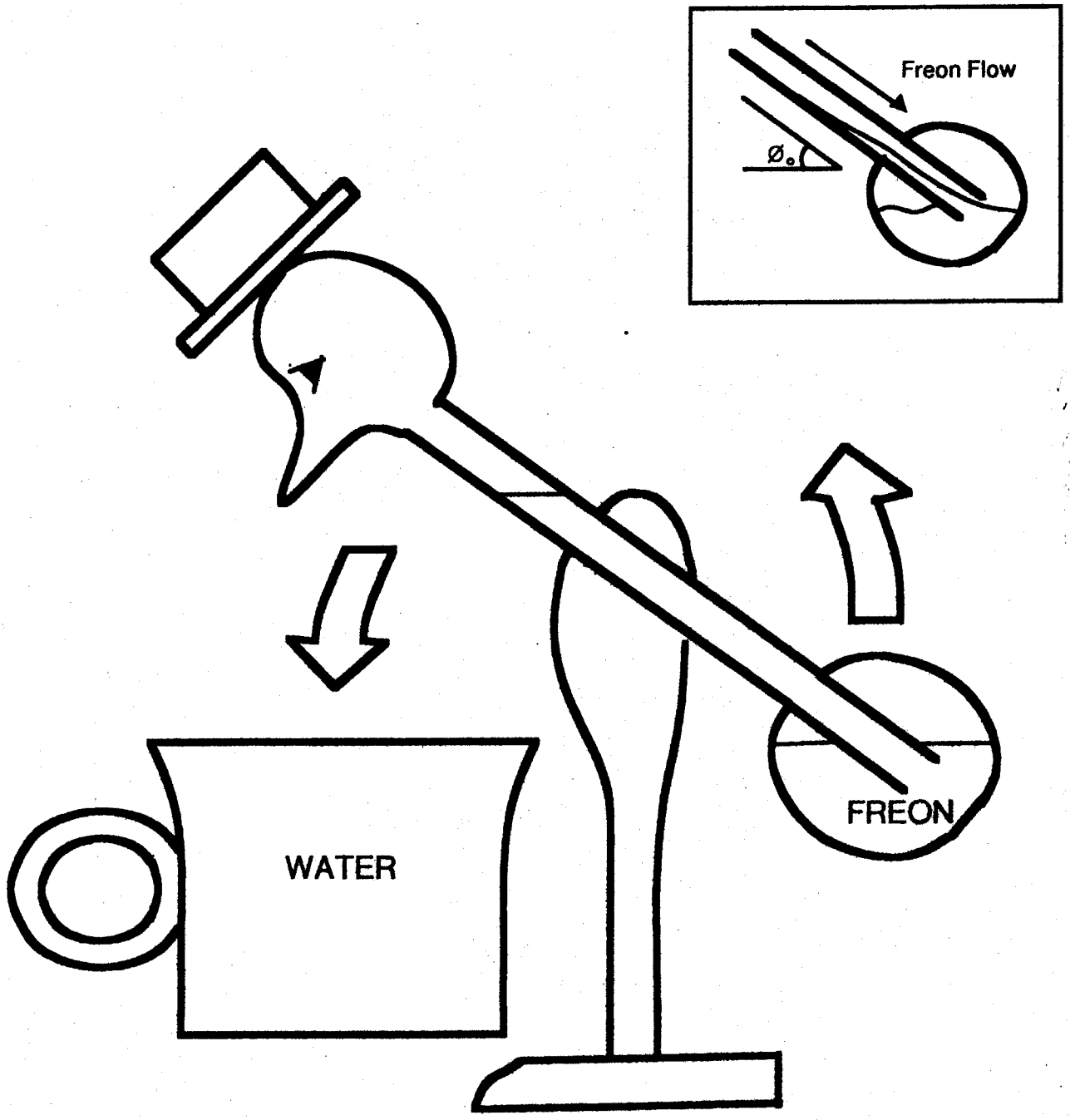
It was mentioned earlier that some descriptions of an object are mutually exclusive. As a basis for discussion of concepts pertaining to this type of horizontal abstraction, let us examine a favorite novelty, the *marvelous drinking bird*. This is intended in part to relieve any monotony, but most important it alludes to the belief that since physiology contains a vast repertoire of causal and associational relationships similar to those in other domains, the computer program being described here must demonstrate its usefulness by applicability to many areas of interest.

The bird is illustrated in figure 6. Initially, the bird's head is submersed in water and then the shaft is set upright. Water evaporation causes the temperature of the gas in the head to decrease. This cooling effect causes gaseous pressure to decrease in the upper bulb. The pressure differential between gas in the upper and lower compartments of the bird causes the freon level to rise. Since the bird is top-heavy, gravity causes the head to slowly fall towards the water to be wetted again.

---

1. *Purpose* and *function* are often confused. "What is the purpose of this mechanism?" is equivalent to asking "Why does this mechanism exist?" Functional information answers the question "What does this mechanism do?".

Fig. 6. The Marvelous Drinking Bird



Let Φ be the angle between the shaft of the bird and the horizontal. As the bird falls, Φ becomes less than some particular value Φ_0 at which gas from the lower bulb may pass upwards into the shaft. This enables gravity to cause freon to flow into the bottom end, changing the bird from a state of being top-heavy to that of being bottom-heavy. When this occurs, gravity causes the bird to slowly become upright and then the cycle is repeated.

Although an ornithologist may object, there are two viewpoints at which to observe the bird's behaviour. When Φ is greater than or equal to Φ_0, we view the level of freon as being determined solely by the pressure difference between the upper and lower bulbs. When Φ is less than Φ_0, we view the level as being determined by gravity causing freon to flow downwards. The two abstract descriptions are at the same level of detail; a change from one to the other corresponds to a horizontal shift of abstraction between two mutually exclusive representations.

In the above example, $\Phi \geq \Phi_0$ and $\Phi < \Phi_0$ were conditions which distinguished between two descriptions of the same object. We will use the term *activation conditions* to describe the circumstances in which a particular viewpoint of an object is valid.

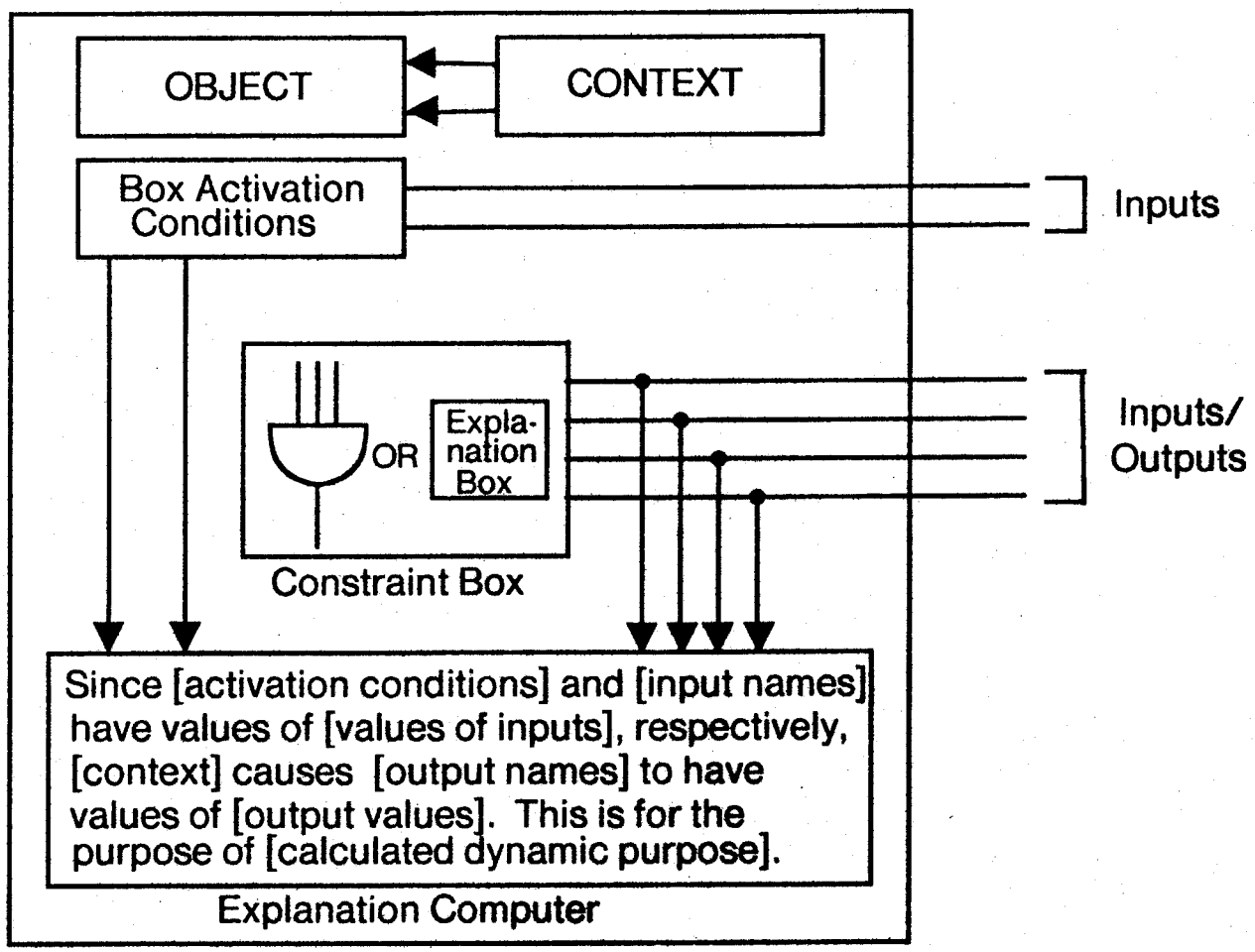
4.4 The Explanation Box

It has been stressed that descriptions of causal contexts should contain information corresponding to the conditions under which the context is valid, the computation performed within the context and the purpose of the computation. The computational device we will use to contain this information is called an *explanation box*. Figure 7 illustrates this construct. The components of an explanation box are described in turn:

1. The *object* specifier describes what abstract or concrete entity we are trying to represent. For example, the object could be "glomerulus".

2. The *context* specifier indicates which viewpoint is being modelled. If the object is "glomerulus", the context could be "a leaky glomerular basement membrane". If the object is "marvelous drinking bird", the context could be "pressure differential determining freon level".

Fig. 7. The Explanation Box



3. *Activation conditions* answers the question "Under what circumstances should the viewpoint described by the explanation box be considered valid?". If the activation conditions of some explanation box are satisfied, that box is considered *active*.

4. As was mentioned previously, for a particular causal context we can specify input/output behaviour. The *constraint box* performs this task. Figure 7 illustrates that within this component, computations can be performed or pointers to explanation boxes at a finer level of description can be found.

5. The *Explanation Computer* contains a framework for explanation creation. It uses information contained in other components of the explanation box. The terms "input pins" and "output pins" are not used because we will not know the manner in which a pin is employed until constraint propagation has ended.

Examples will be used later to solidify these concepts.[1]

4.5 Explanation of Constraint Propagation

After all input information has been propagated as far as possible within the constraint network, NEPHROS enters the explanation phase. To allow examination of causal behaviour within a particular physiological context, the user types in the name of the corresponding explanation box. The program then asks which pin of the constraint box is to be examined. For example, if the user chooses to examine the "kidney", "sodium excretion" may be the pin name given as input. If the pin has a value, NEPHROS will begin a sequence of instructions to explain why. Before going further, we should make clear in our minds the mechanisms which cause a pin to be assigned a value:

1. The user can directly set the pin in the input phase. For example, the user may decide to set "cardiac output" of the "heart" to *decreased*.

---

1. Those familiar with the *frame theory* introduced by Marvin Minsky, may realize I have been trying to present the explanation box as a special type of frame. A more complete discussion is given in the Epilogue.

2. A pin can be given a value by being connected to a pin belonging to some other constraint box. As an illustration, NEPHROS sets up a connection between "ADH Secretion" of the "ADH Complex" explanation box and "ADH Level" of the "Arterial Blood" explanation box before any constraint propagation occurs. If one of these pins is given a value, that value is automatically transferred to the other pin.

3. Computations within the constraint box to which a pin belongs may cause that pin to be assigned a value.

Number 3 needs further elaboration. It was mentioned earlier that within a constraint box there may be true constraints, such as an IQ-adder, or pointers to explanation boxes which describe causality at a finer level of detail. In order to explain why a particular pin has a value, we may have to go down several layers of abstraction until we find constraint boxes with "substance". Therefore, explanation of pin values is recursively defined.

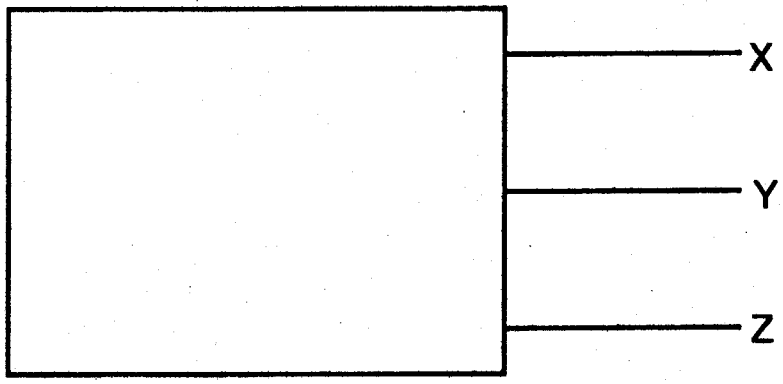
Figure 8 will be used to illustrate the constraint propagation and explanation issue. The upper figure represents a black box which assigns to z the value of (x − 1) times y. The lower figure is a grey box showing the internal workings of the black box. We can see that z is assigned the value of xy minus y. The symbol = is used to represent pin *equivalence*. Equivalence exists between two pins when one has a value if and only if the second has the same value. For this reason, pins x^{I} and y^{I} are immediately assigned the values of x and y, respectively. Pins y^{II} and z^{II} are given the values of pins y and z, respectively.

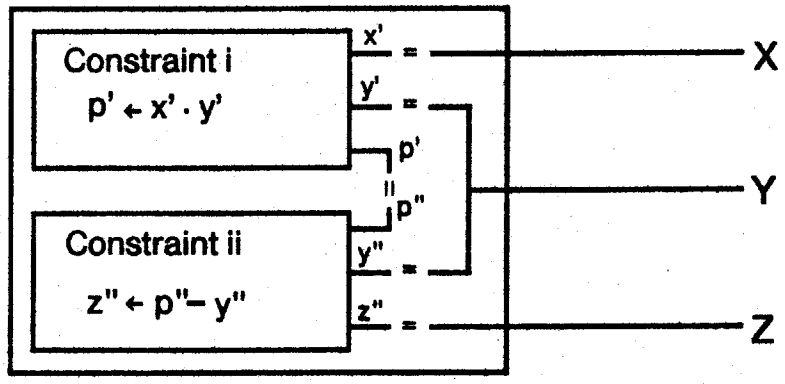
Let us set x to 3 and y to 2. Constraint i assigns to pin p^{I} the product of x and y. Since pin p^{I} of constraint i is connected to pin p^{II} of constraint ii, the value of 6 is propagated from pin p^{I} to pin p^{II}. Constraint ii calculates the difference between the values of pin p^{II} and pin y^{II}. Since pin y^{II} has the value 2, z is computed to be 4.

Now let us ask ourselves "Why does z have the value 4?." If we are allowed to use the grey box, the following rationalization may be given (a → b means "b was assigned the value a"):

Since	$\{4 \rightarrow z^{II}\}$,	equivalence	caused	$\{4 \rightarrow z\}$
∵	$\{6 \rightarrow p^{II}, 2 \rightarrow y^{II}\}$,	constraint ii	caused	$\{4 \rightarrow z^{II}\}$
∵	$\{2 \rightarrow y\}$,	equivalence	caused	$\{2 \rightarrow y^{II}\}$

Fig. 8. Why Does Z Have Its Value?





$\because$ $\{6 \rightarrow p^{I}\}$, equivalence caused $\{6 \rightarrow p^{II}\}$

$\because$ $\{3 \rightarrow x^{I},\ 2 \rightarrow y^{I}\}$, constraint i caused $\{6 \rightarrow p^{I}\}$

$\because$ $\{3 \rightarrow x\}$, equivalence caused $\{3 \rightarrow x^{I}\}$

$\because$ $\{2 \rightarrow y\}$, equivalence caused $\{2 \rightarrow y^{I}\}$

Some of the equivalence links may seem to be unnecessary as they make explanation rather cumbersome. It should be emphasized that equivalence links are created to ensure that every pin belongs to one and only one constraint box. This allows clean separation of explanation boxes at the same level of abstraction and allows distinction of different levels of vertical abstraction. For example, a link placed between p^{I} and p^{II} separates constraints i and ii. A link placed between x^{I} and x allows the pin entering constraint i to be distinguishable from the pin entering the black box, which is at a higher level of abstraction. To avoid explicit reference to equivalence links which provide no useful information to the explanation phase of NEPHROS, equivalences are flagged internally as useful for explanatory purposes or not.

If given only the *black* box, how do we explain why z has the value 4? The following explanation may be given: since $3 \rightarrow x$, $2 \rightarrow y$ and the black box computes $z = (x - 1)\,y$, z is assigned the value 4. However, for black boxes with many pins and many functions, compact summaries of lower level behaviour such as "$(x - 1)\,y$" usually cannot be found. When faced with the task of explaining why a particular pin of a large black box received its value, we cannot even say for sure that all other pins of the box were involved in the computation of that value! Therefore, the task at hand is to determine how we can send useful information from the grey box level up to the black box level.

4.6 The Causality Relation C

When describing computation in the grey box of figure 8, we made statements such as the following: "since $3 \rightarrow x^{I}$ and $2 \rightarrow y^{I}$, constraint i caused p^{I} to be 6." In some sense, x^{I} and y^{I} are responsible for p^{I} being assigned a value. To capture this concept, the *Causality Relation*, **C**, is defined as follows:

a C b if and only if **a** is (partly) responsible for **b**.

In order to graphically represent the direction of causality, **a** $\xrightarrow{\text{causes}}$ **b** will be used instead of

a C b. With respect to the computation of p^l described above, we have $\{3 \rightarrow x^l\} -_{causes} \rightarrow \{6 \rightarrow p^l\}$ and $\{2 \rightarrow y^l\} -_{causes} \rightarrow \{6 \rightarrow p^l\}$.

For our purposes, the relation C must be specifically tailored to describe envisionment of processes occurring in a physiological system. We will use the term *event* to describe the situation in which a physiological variable is assigned a value. For example, the assignment of *decreased* to "cardiac output" of the "heart" is an event. Since events in an envisionment correspond to real word occurrences, the causality relation allows us to express the fact that one physiological happening was involved in causing another to occur.

We must take cognizance of the fact that the NEPHROS physiological network contains multidirectional constraints. For example, if water reabsorption in the collecting tubule of the kidney is *increased*, the program concludes that ADH level in the blood must be *increased*. In this instance, information propagation is in the direction opposite to *physiological* causality. To dispel any confusion, we shall use the statement "**a** $-_{causes} \rightarrow$ **b**" to mean "due to constraint propagation, event **a** is partly responsible for event **b**". This may or may not be in the normal direction of physiological occurrences.

The causality relation C is a tool we can use to describe physiological processes. This relation has several important properties, which are listed below:

1. Let **a** and **b** be two physiological events such that **a** $-_{causes} \rightarrow$ **b**. Since **a** and **b** correspond to events during constraint propagation, **b** occurs later than **a** in the order of events of the propagation. Therefore, we could never have the case that **a** $-_{causes} \rightarrow$ **a**. One may argue that in cases of positive feedback, one physiological perturbation may indirectly cause the same perturbation.[1] However, these two events are distinguishable since they are at different points in the ordering of events occurring during information propagation.

2. Assume we know that **a** $-_{causes} \rightarrow$ **b**. Using the same reasoning as in 1., we could never have the case that **b** $-_{causes} \rightarrow$ **a**.

---

1. For example, a decreased output of blood from the heart reduces blood flow to the tissues. If the blood flow to the heart tissue itself is not adequate, cardiac output worsens.

3. Note that the definition for C states that **a** $-_{causes}\rightarrow$ **b** if the value assignment labelled a partly causes the value assignment **b** to occur. This does not mean that a has to be directly involved in the computation. In the grey box example discussed earlier, since $\{3 \rightarrow x\} -_{causes}\rightarrow \{3 \rightarrow x'\}$ and $\{3 \rightarrow x'\} -_{causes}\rightarrow \{6 \rightarrow p'\}$, we can say that $\{3 \rightarrow x\} -_{causes}\rightarrow \{6 \rightarrow p'\}$.

These properties will allow us to put a mathematical label on the relation C.

4.7 C As A Partial Ordering

Let us consider the three characteristics of the physiological causality relation described in the last section. In set theoretic terms, these properties correspond to irreflexivity, antisymmetry and transitivity, respectively. This permits C to qualify as a *strict partial ordering* on the set of physiological events:

Definition[1] R is a *strict partial ordering* on X if and only if

(i) $\forall x \in X [\neg (xRx)]$ (irreflexivity)

(ii) $\forall x, y \in X [xRy \Rightarrow \neg(yRx)]$ (antisymmetry)

(iii) $\forall x, y, z \in X$ [xRy and yRz $\Rightarrow$ xRz] (transitivity)

To enable us to have a concrete sense of this abstract concept, consider the following scenario:

> Two friends G and H cross paths one afternoon, exchange salutations and decide to meet at 8 pm in the evening to construct a physiological constraint network. Contemplating success, G decides they should both prepare for a celebration and volunteers to pick up food from store 1 and then beverages from store 2. After the two depart, H realizes he has nothing to do until 8 pm and decides to have an afternoon swim.

Let T be a relation such that aT b means "a occurs before b". For example, if x is the event "G is at store 1" and y is the event "G is at store 2", we have xTy. The reader can easily verify that T is a strict partial ordering over the events in our story.

---

1. From [27]. The reader will note that rule i is redundant. If we assume, per absurdum, there is an x such that xRx, rule ii will permit the coexistence of xRx and $\neg$(xRx), clearly a contradiction.

The relation T allows us to order the events occurring in G's day in a linear manner. The same is true for friend H. However, we cannot tell whether the event "H takes a swim" occurs before or after "G is at store 1". It is also not known how "H takes a swim" and "G is at store 2" are related. This is the most important characteristic of a partial ordering, it gives us the power to order certain entities in a linear fashion, but gives us the flexibility to leave unspecified the ordering between other entities.

4.8 C^{-1} As A Partial Ordering

Relation C has been tagged as a partial ordering over physiological events. Although interesting, the mathematical label "partial ordering" provides us with no useful information unless some property of partial orderings is directly applicable to the constraint propagation and explanation issue. Consider the definition and corresponding theorem below:

Definition Let R be a relation. The relation R^{-1} (pronounced "R inverse") is defined by:

$xR^{-1}y$ if and only if yRx.

Theorem If R is a strict partial ordering on set X, R^{-1} is also a strict partial ordering on set X.

Proof of theorem See Appendix 2.

In the grey box example given earlier, we traversed causal pathways in the reverse direction to explain why z was assigned the value 4. Since C is used to express relationships of the form $a \xrightarrow{\text{causes}} b$, C^{-1} describes relationships such as $b \xrightarrow{\text{caused-by}} a$. Therefore, to explain why a certain physiological variable has a value, we use C^{-1} to travel backwards over causal pathways created during forward propagation. Since C is a partial ordering on the set of physiological events, the theorem above tells us that C^{-1} is also a partial ordering on that set. In fact, if we look closely at the proof, C^{-1} is the *same* partial ordering with only the directionality changed.

The previous discussion indicates that our explanation of physiological constraint propagation has an underlying mathematical structure that is as "strong" as the constraint propagation itself.[1] Therefore, the explanation phase will provide as much information as is derived by constraint propagation in the forward direction.

4.9 Creating A Pedagogical Network

Explanation boxes have been introduced as devices which will allow us to derive explanations of physiological processes. For each explanation box there corresponds exactly one constraint box which carries out the computations belonging to that causal viewpoint. We shall see in the next chapter that by making equivalent the appropriate constraint pins, we can build a network of explanation boxes representing horizontal and vertical abstraction.

Explanation boxes were designed to allow us to build a hierarchy of descriptive elements. However, for a given explanation box we may have to go down several levels of abstraction before *real* computations are found. This should not alarm us since we do have considerable descriptive power by having such a hierarchy. The question that is as yet unanswered is how we can push information pertaining to lower level constraint propagation to higher levels of abstraction. The transitive property of the causality relation C will allow this to happen. After constraint propagation has finished, we will know all instances of $a \xrightarrow{\text{causes}} b$. By the transitive property of C, if $a \xrightarrow{\text{causes}} b$ and $b \xrightarrow{\text{causes}} c$, we know that $a \xrightarrow{\text{causes}} c$ *indirectly*. The set of all such relationships is the transitive closure of the causality relation C. The closure contains elements of the form $a \xrightarrow{\text{causes}} b$, which may summarize several steps of causation.

NEPHROS creates the causal closure after all information has been propagated. The explanation phase is then entered in which an explanation box is chosen for examination. When the value of a particular constraint pin of that box is to be justified, NEPHROS refers to the causal closure to determine which other pins belonging to the explanation box are directly or indirectly responsible for the pin's value assignment. This permits us to build explanations at the black box

---

1. When reversing a directed network, we are not always so lucky. For example, the relation "consists of" forms a tree. The inverse relation , "is an element of", intuitively does not carry as much information and forms a digraph in which pathways converge.

level.

As shown earlier, the program user chooses a context in which to examine causal behaviour. Commands may be typed into NEPHROS which allow passage both horizontally and vertically within the hierarchy of explanation boxes representing renal physiology. If SAME LEVEL is typed in, justification for a particular pin's value is given at the horizontal level of abstraction the user has finally chosen. By passing among explanation boxes at the same level, NEPHROS traces back causal pathways which were created by constraint propagation.

With limitless energy, NEPHROS will expound upon any physiological information within its descriptive framework. We shall use the term *pedagogical network* to describe the program's control structure and hierarchy of explanatory devices.

5. A Pedagogical Network For Renal Physiology

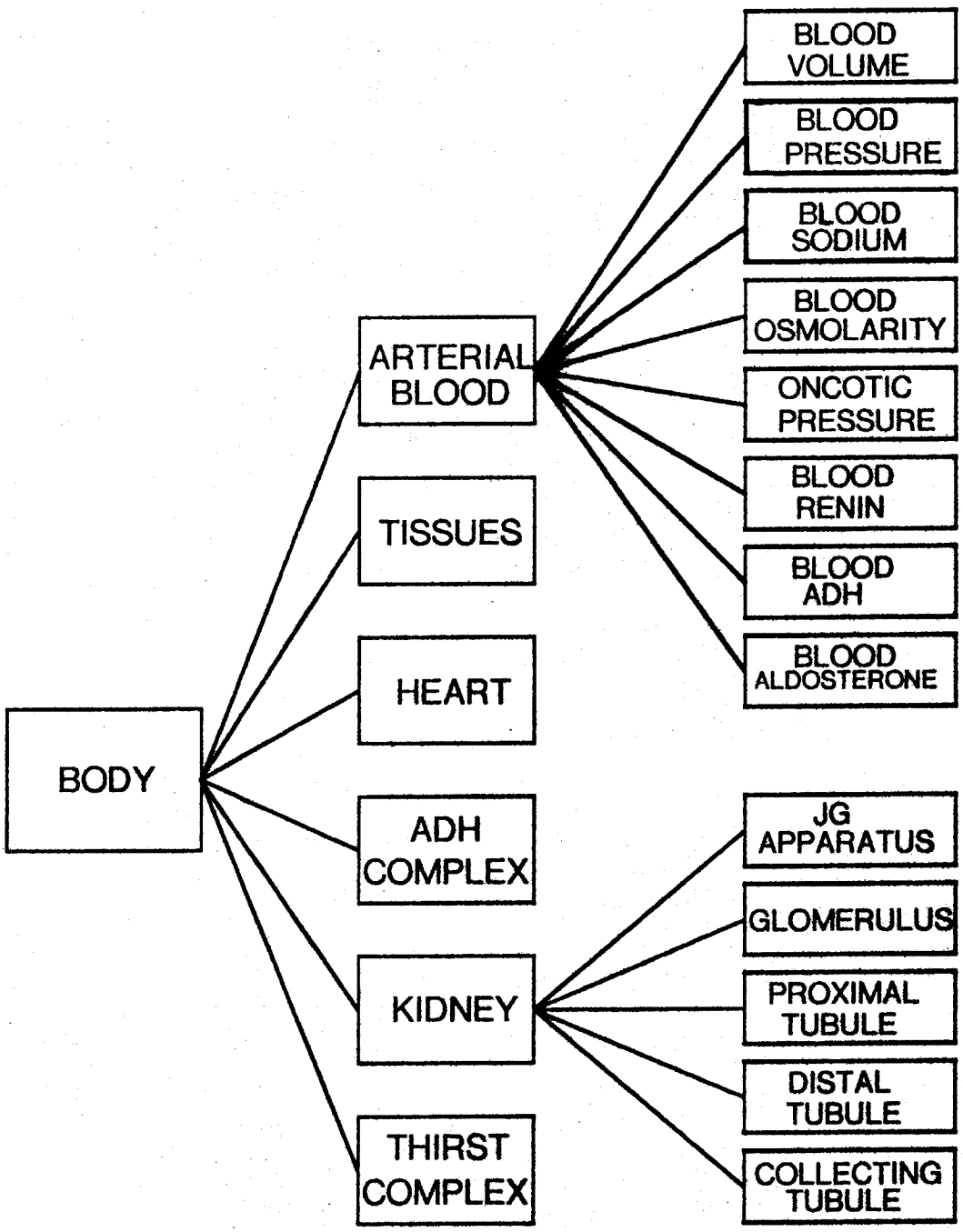
5.1 The Explanation Box Hierarchy

In this chapter, the NEPHROS hierarchy of physiological explanation boxes shall be described. The underlying constraint network models the subset of renal physiology pertaining to hypoperfusion of the kidney, nephrotic syndrome and the syndrome of inappropriate secretion of antidiuretic hormone. Kenneth Forbus of M.I.T.'s Artificial Intelligence Laboratory created the 1980 version of the Conlan Constraint Language upon which the NEPHROS network of causal relationships was built[8].

The hierarchy of explanatory devices used by NEPHROS is shown in Figure 9. The labels in the diagram are *object specifiers* of the corresponding explanation boxes. Each box can be considered as a subcontext within the causal context of its "parent." For example, "Collecting Tubule" is an object we refer to when discussing the "Kidney". It should be noted that some explanation boxes in the figure refer to concepts which are not anatomically concrete. For example, "Blood Osmolarity" is not tangible in the classical physical sense.

From the time the computer program begins executing until it has completed its task, the structure shown in figure 9 remains fixed. The boxes in the diagram correspond to objects which are considered to permanently exist in the body. When pathological processes are present, some of these objects may be diseased and we cannot use the *default* explanation boxes which describe normal behaviour. For each object shown in the figure, the program must decide which one of several mutually exclusive causal viewpoints is to be implemented in the network. *Activation conditions* permit this to be done.

Fig. 9. The Explanation Box Hierarchy



5.2 The Body

Let us focus our attention on the manner in which renal physiological relationships are modelled within NEPHROS. To enable this, certain physiological systems shall be described in full detail. So that the reader unfamiliar with medicine will be put at ease, this presentation will not be flowered with medical jargon unimportant to the points at hand. A glossary defining important terms is provided as an appendix; the first occurrence of all words defined in the glossary is underlined in the text.

Figure 10 illustrates schematically how NEPHROS views the body. The reader will recognize the main circulatory loop which consists of the heart, the blood vessels and the tissues. Only the arterial component of the circulatory system is labelled in the diagram. The three syndromes being modelled did not require an explicit representation of the venous system. Since the ADH complex, thirst complex and kidney play special roles in renal physiology, separate explanation boxes have been created for them. All other body parts are relegated to the Tissues explanation box. Arrows indicate directions in which information is allowed to propagate.

The term homeostasis is used to describe the condition in which the body is internally balanced. The ADH complex, thirst complex and kidney help to keep the body in this steady state by being regulators of water and electrolytes. They can be viewed as sensor/effector mechanisms which derive information from the blood and return substance or additional information back to the blood. The explanation boxes corresponding to these regulators are directly connected to the Arterial Blood explanation box. To avoid unnecessarily long pathways of envisionment, the three water and electrolyte regulators have not been included within the main circulatory loop.

In figure 11, the Body explanation box is shown in greater detail. Pins connected vertically are considered equivalent. A legend for the diagram is given on the following page. The reader will notice that the activation conditions are described as DEFAULT. Any explanation boxes labelled in that manner are to be used if no other boxes modelling the same object are active. Since NEPHROS has only one viewpoint of the body as a whole, the explanation box shown in figure 11 will always be active. One may argue that there should be a similar causal context for each disease within the domain of renal physiology. However, it should be emphasized that the Body explanation box only postulates the existence of the body's substructures and their

Fig. 10. The Body

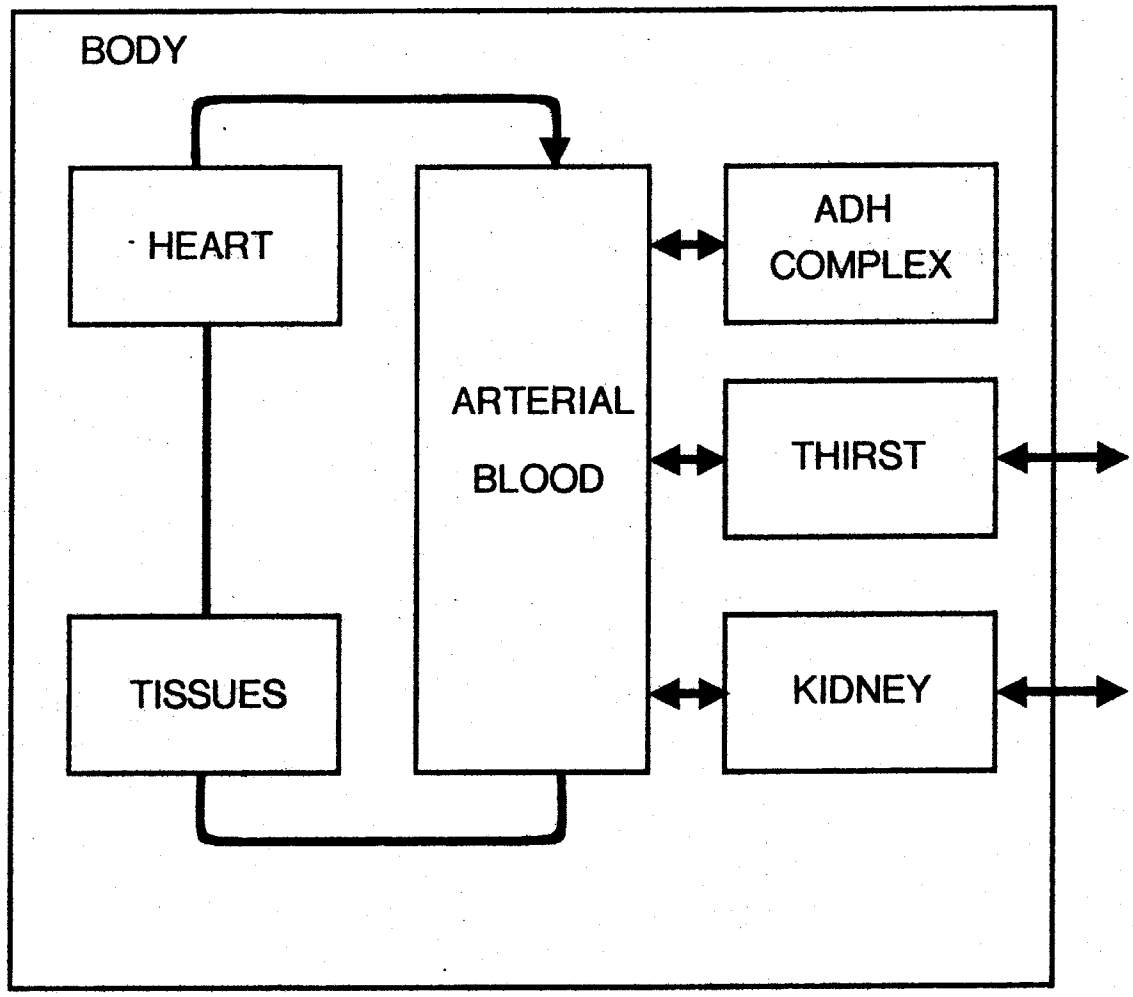


Fig. 11. The Body Explanation Box

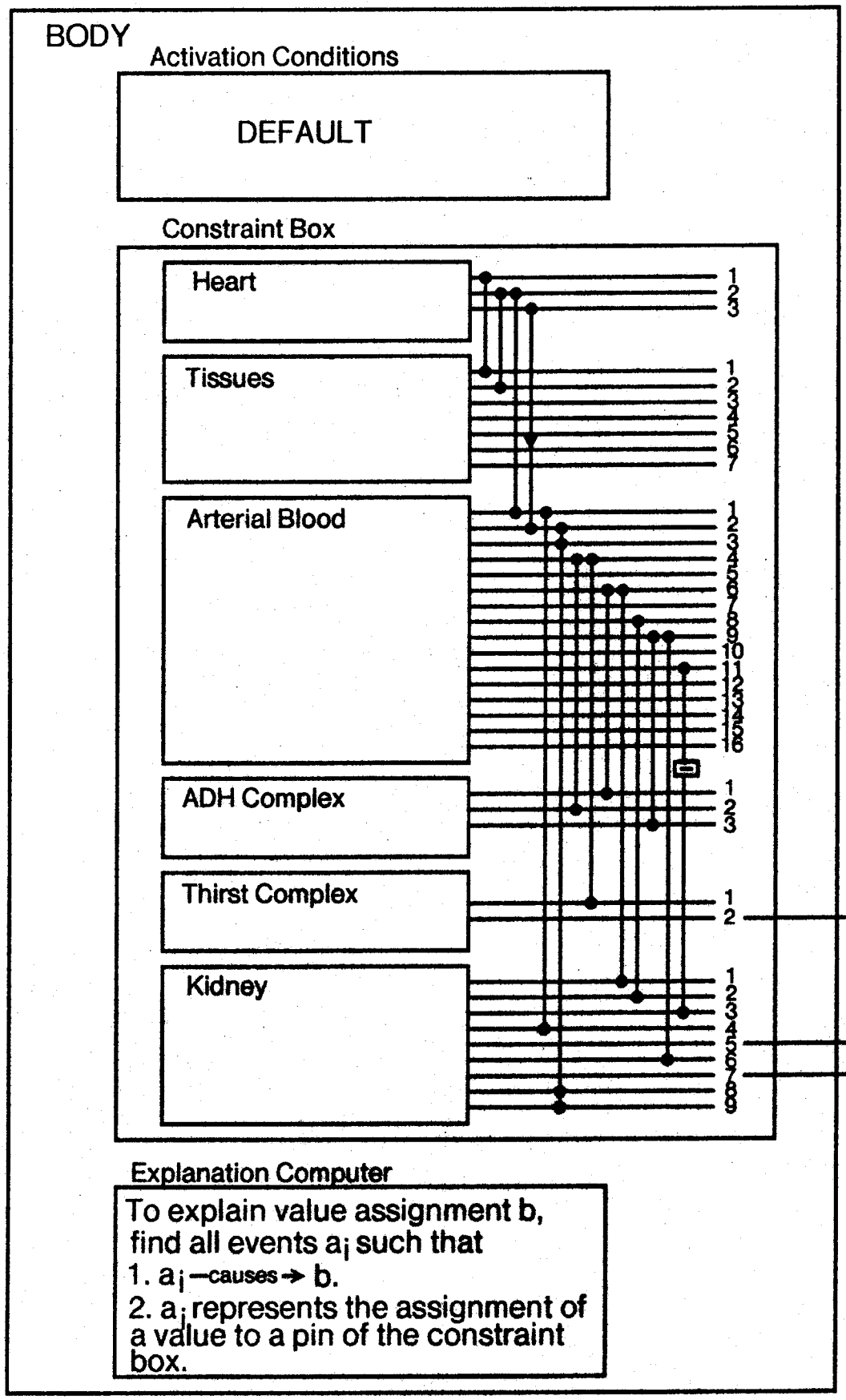


Fig. 12. Legend for Body Explanation Box

CONSTRAINT		PIN NAME
BODY	1	WATER INTAKE
	2	KIDNEY SODIUM EXCRETION
	3	KIDNEY WATER EXCRETION
HEART	1	VENOUS RETURN
	2	ALDOSTERONE LEVEL
	3	CARDIAC OUTPUT
TISSUES	1	VENOUS RETURN
	2	ALDOSTERONE PRODUCTION
	3	RENIN LEVEL
	4	BLOOD VOLUME
	5	CAPILLARY HYDROSTATIC PRESS.
	6	CAPILLARY ONCOTIC PRESSURE
	7	INTERSTITIAL FLUID VOLUME
ARTERIAL BLOOD	1	ALDOSTERONE LEVEL
	2	CARDIAC OUTPUT
	3	EFFECTIVE BLOOD VOLUME
	4	BLOOD OSMOLARITY
	5	TOTAL PERIPHERAL RESISTANCE
	6	BLOOD PRESSURE
	7	BLOOD SODIUM
	8	BLOOD RENIN
	9	ADH LEVEL
	10	EFFECTIVE BLOOD FLOW
	11	BLOOD PROTEIN
	12	BLOOD ONCOTIC PRESSURE
	13	KIDNEY RENIN PRODUCTION
	14	ADH PRODUCTION
	15	KIDNEY SODIUM RETENTION
	16	HEART ALDOSTERONE LEVEL
ADH COMPLEX	1	BLOOD PRESSURE
	2	BLOOD OSMOLARITY
	3	ADH SECRETION
THIRST	1	BLOOD OSMOLARITY
	2	WATER INTAKE
KIDNEY	1	BLOOD PRESSURE
	2	RENIN PRODUCTION
	3	PROTEIN EXCRETION
	4	ALDOSTERONE LEVEL
	5	SODIUM EXCRETION
	6	ADH LEVEL
	7	WATER EXCRETION
	8	BLOOD FLOW
	9	EFFECTIVE BLOOD VOLUME

interconnections and not their pathological states.

Before leaving the Body explanation box, a word should be said about the explanation computer. Although the constraint box contains only three external pins, the general algorithm for explanation derivation at the black box level is shown. If we wish to justify event **b**, the value assignment of a particular pin of the constraint box under consideration, NEPHROS searches the transitive closure of the causality relation C for all relationships of the form a_i – causes → **b**, such that event a_i represents the value assignment to a pin of the constraint box. The explanation computer uses all such a_i in the description of causal behaviour.

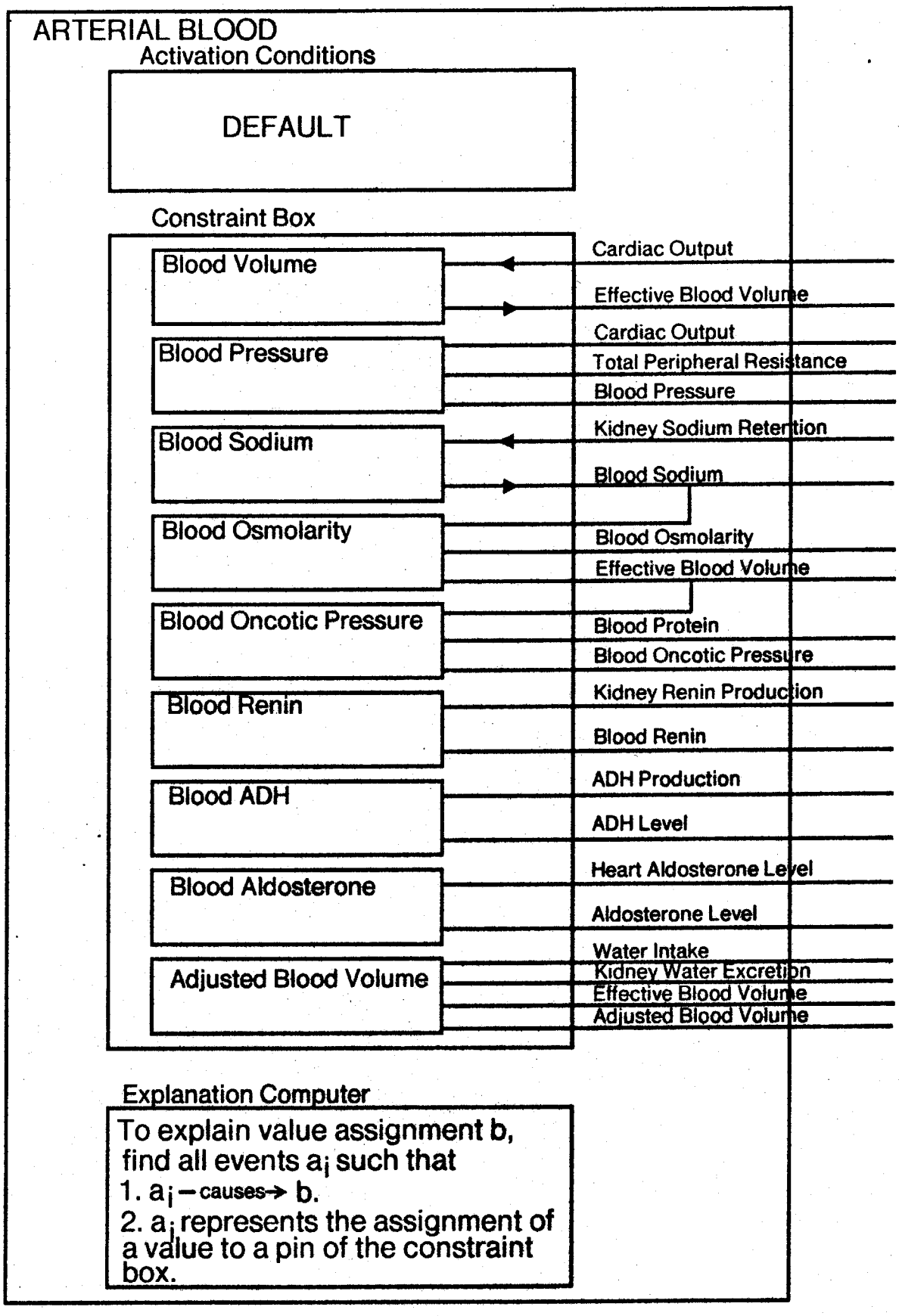
5.3 Heart

The Heart explanation box is presented in the next chapter.

5.4 Arterial Blood

The Arterial Blood explanation box is shown in figure 13. Boxes contained within the constraint box are properties of the blood. Effective blood volume, blood pressure, blood osmolarity and blood oncotic pressure are abstract, though measurable, properties of the blood. The explanation boxes describing sodium, renin, ADH and aldosterone refer to absolute levels of these substances as opposed to concentrations. For each causal context within the constraint box, there is a corresponding explanation box with activation conditions, constraint box and explanation computer. In order that we not drown ourselves in detail at this point, descriptions of these computational elements are relegated to appendix 3.

Fig. 13. The Arterial Blood Explanation Box



5.5 ADH Complex

To maintain homeostasis, the body must have the ability to retain water during periods of fluid loss or deprivation. The collecting tubule of the kidney is one of the important sites for fluid regulation. An increased level of antidiuretic hormone (ADH) at that location will cause water reabsorption to increase.

The control center for ADH secretion is located in the hypothalamus of the brain. An increased blood osmolarity or a decreased arterial blood pressure will cause the ADH complex to increase production of the hormone.[1] An explanation box describing the ADH complex is shown in Figure 14.

5.6 Thirst Complex

The thirst complex, also located in the hypothalamus, provides an additional mechanism for water regulation. An increased blood osmolarity will cause thirst to increase. The thirst complex is illustrated in figure 15.

5.7 Kidney

A great deal of research has been done in the area of renal physiology, and many details of the intricate workings of the kidney have been uncovered. Pathological disorders have been described in terms of the minute functional subunit of the kidney, the nephron.[2] However, as in other medical research areas, some facts are isolated and more information about this homeostatic marvel is yet to be discovered. Here we shall assume a fixed set of renal physiological principles to be true.

---

1. More correctly it is extracellular fluid osmolarity, rather than blood osmolarity, which causes an increased ADH production.
2. Although there are two kidneys, a one-kidney model will be sufficient to represent the physiology underlying the syndromes described in the next chapter. And as in most physiology texts, we will combine the kidney's million nephrons into one.

Fig. 14. The ADH Complex Explanation Box

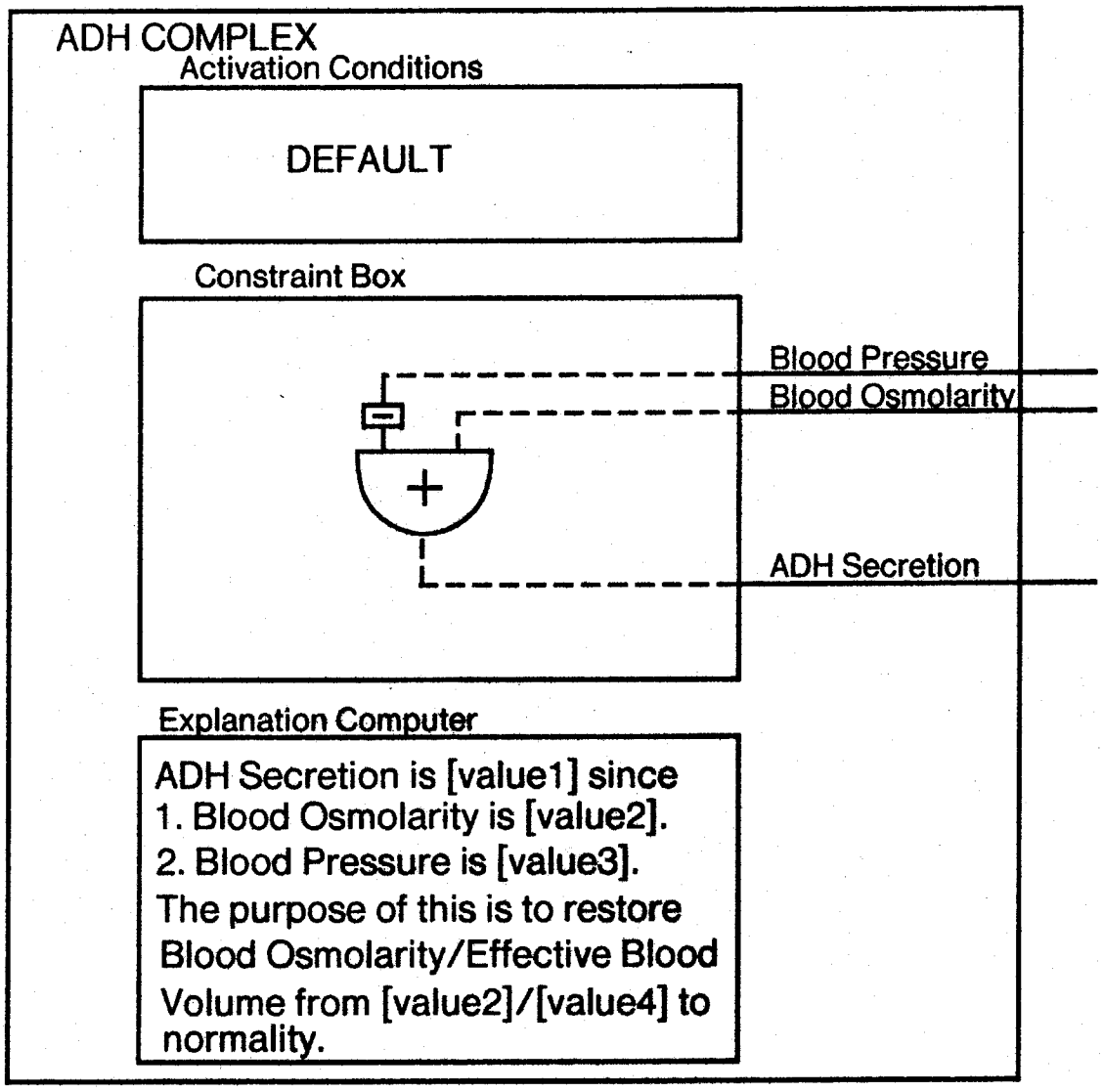


Fig. 15. The Thirst Explanation Box

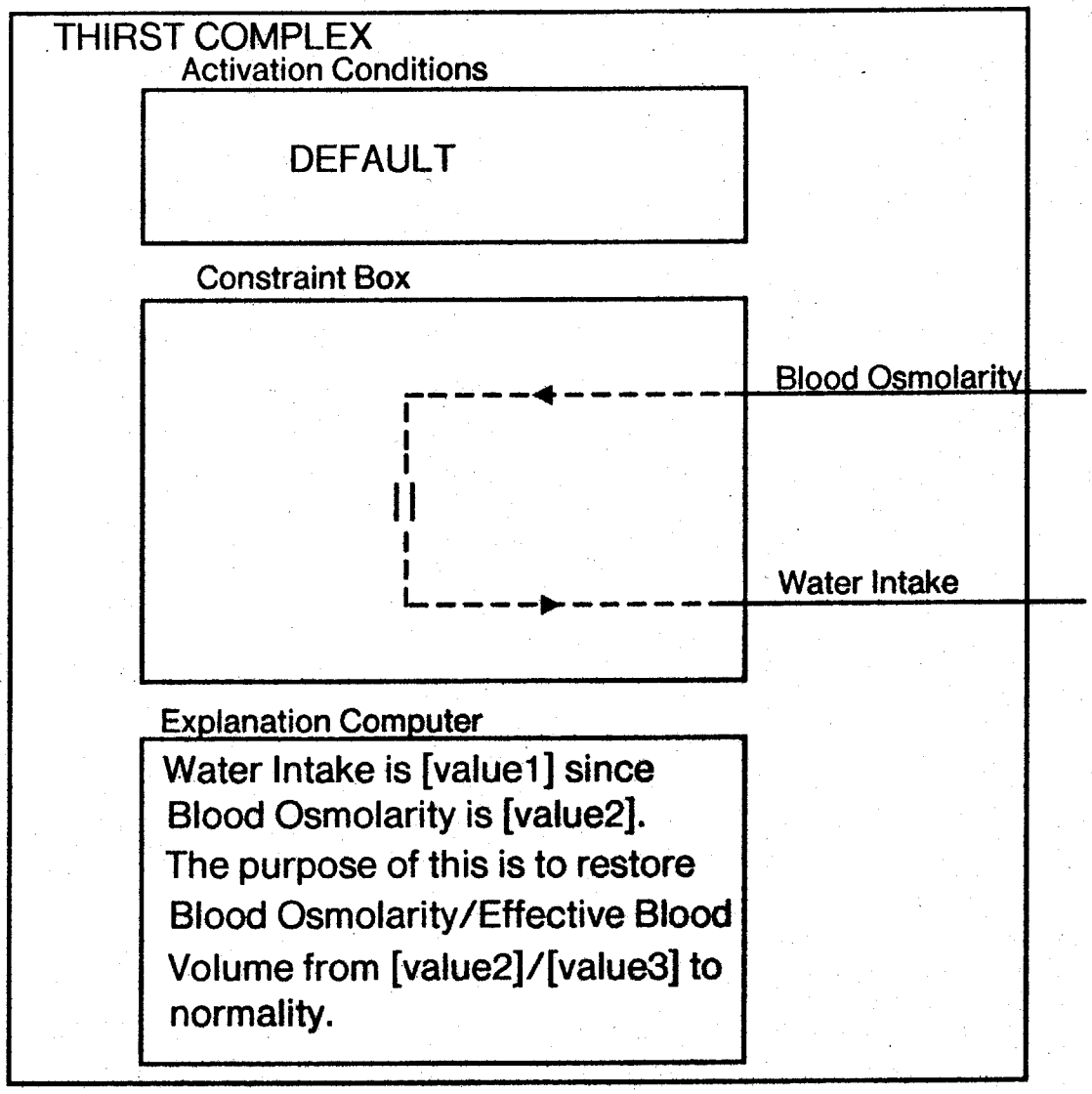
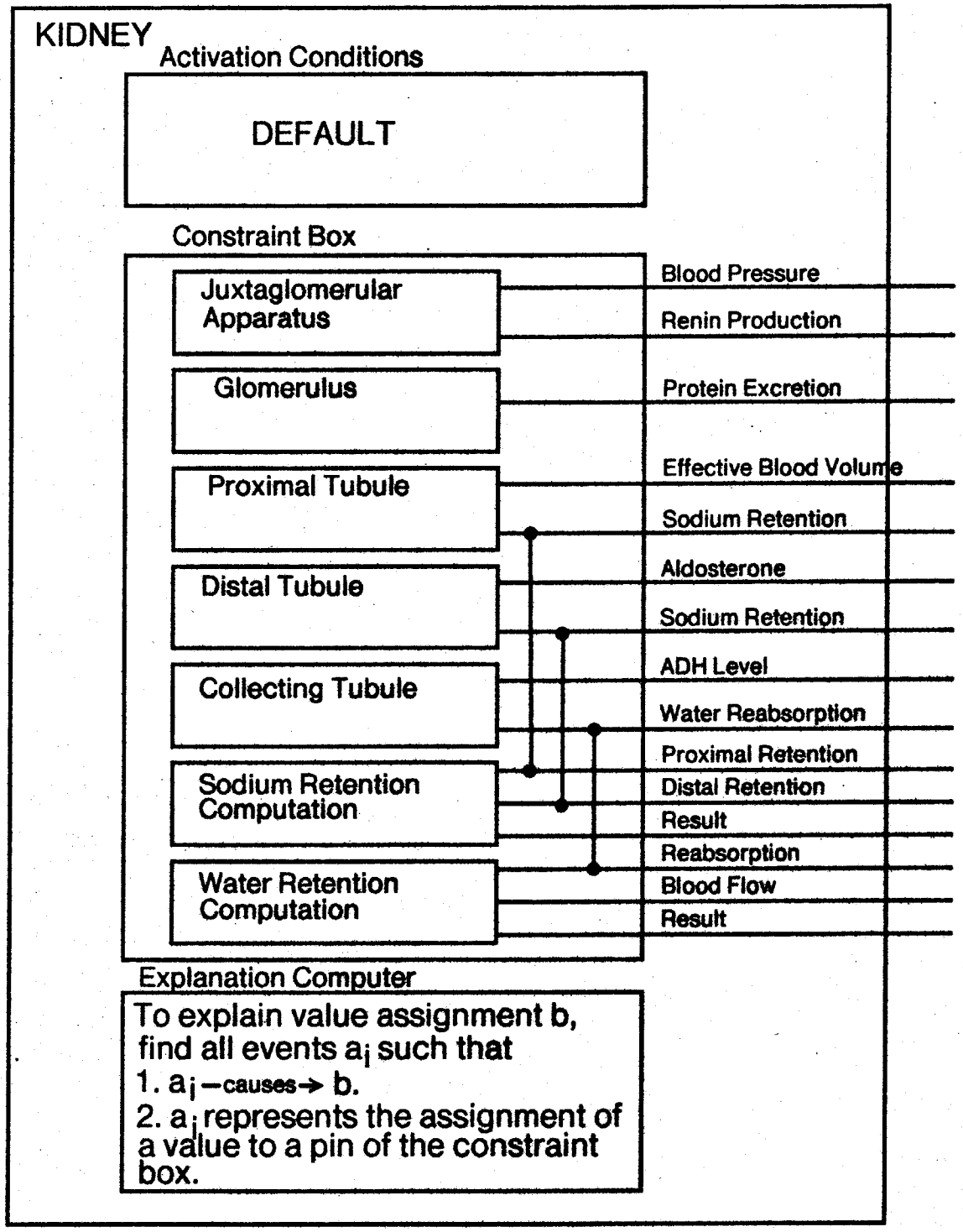


Fig. 16. The Kidney Explanation Box



The Kidney explanation box is shown in figure 16. Within the constraint box, the nephron is decomposed into the JG apparatus, glomerulus, proximal tubule, distal tubule and collecting tubule.

5.8 Juxtaglomerular Apparatus

There are pressure receptors in the juxtaglomerular(JG) apparatus of the nephron which can sense a change in blood pressure. The JG apparatus responds to a decreased pressure by increasing its production of renin. Similarly, an increase in blood pressure will cause a decrease in renin production. In the blood, renin acts as an enzyme to convert a compound called renin substrate into angiotensin I. An enzyme in the lungs converts angiotensin I into angiotensin II. Angiotensin II stimulates the adrenal cortex to produce aldosterone. The last three steps may be summarized as "the tissues convert renin into aldosterone".

By the mechanism described above, changes in blood pressure cause appropriate changes in the level of blood aldosterone. Aldosterone acts within the kidney to increase sodium retention and thus increase blood osmolarity. Since both the ADH complex and the thirst complex are sensitive to changes in blood osmolarity, they attempt to restore osmolarity to normal by increasing water retention and intake, respectively. Therefore, the renin-angiotensin-aldosterone system indirectly causes water reabsorption by increasing sodium reabsorption.

The explanation box in figure 17 is a model of the juxtaglomerular apparatus.

5.9 Distal Tubule

As discussed above, under certain circumstances the body's tissues will secrete aldosterone to maintain water and electrolyte balance. Aldosterone has its main effect on the distal tubule of the nephron, where it causes reabsorption of sodium ions. The Distal Tubule explanation box is shown in Figure 18.

Fig. 17. The Juxtaglomerular Apparatus Explanation Box

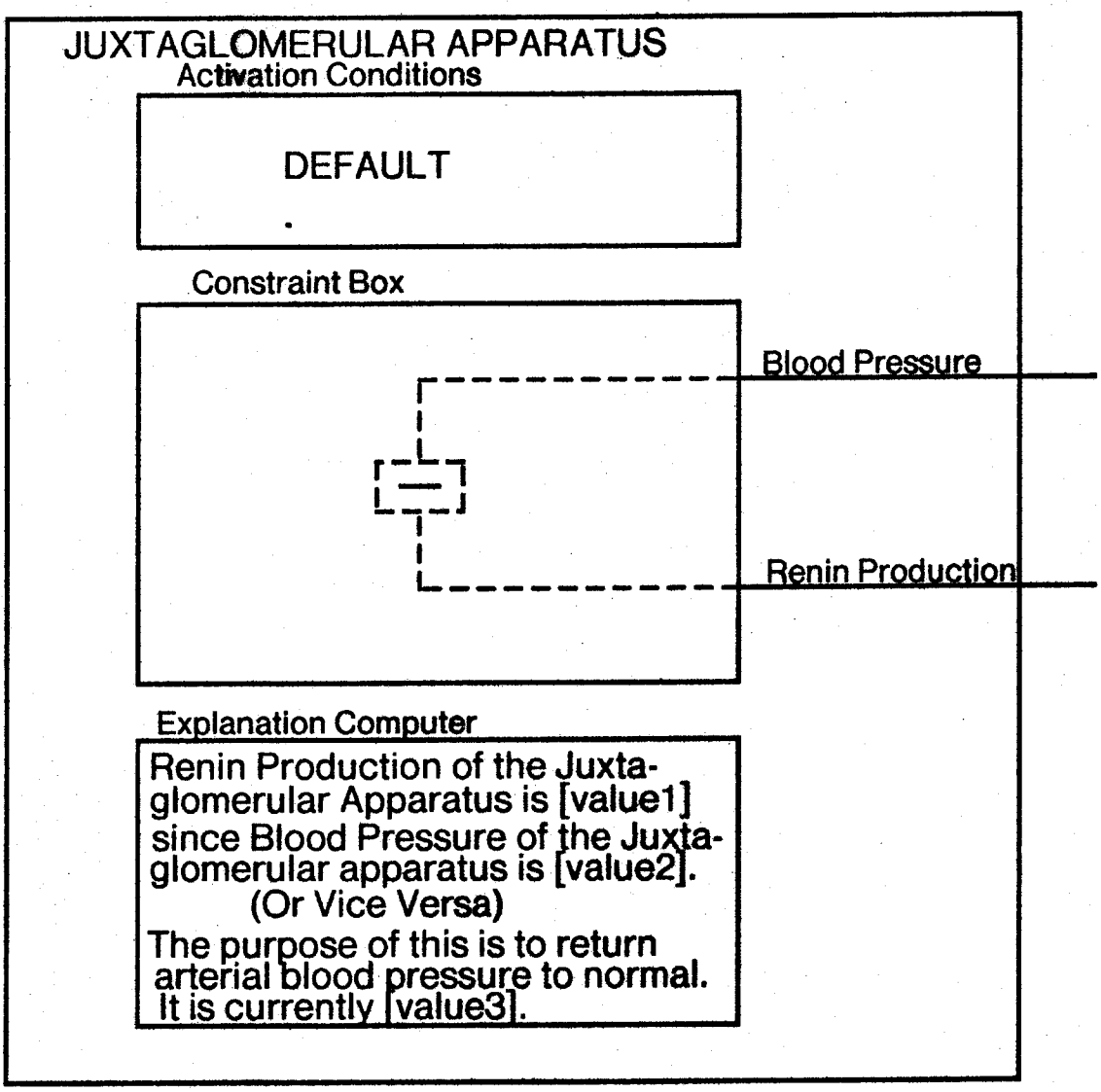
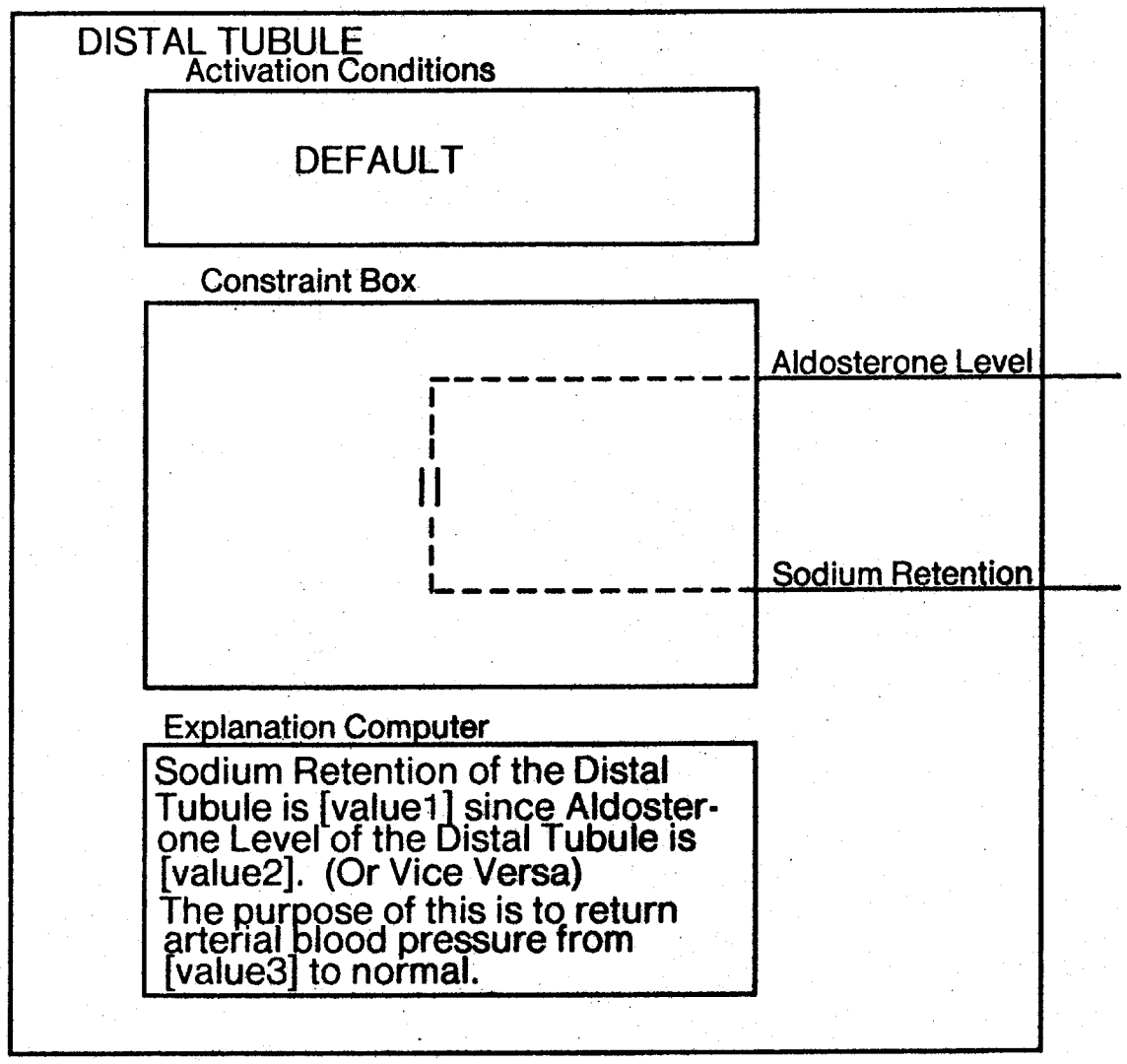


Fig. 18. The Distal Tubule Explanation Box



5.10 Glomerulus

An explanation box describing the glomerulus causal context shall be introduced in the next chapter.

5.11 Proximal Tubule

An explanation box representing the proximal tubule shall be introduced in the next chapter.

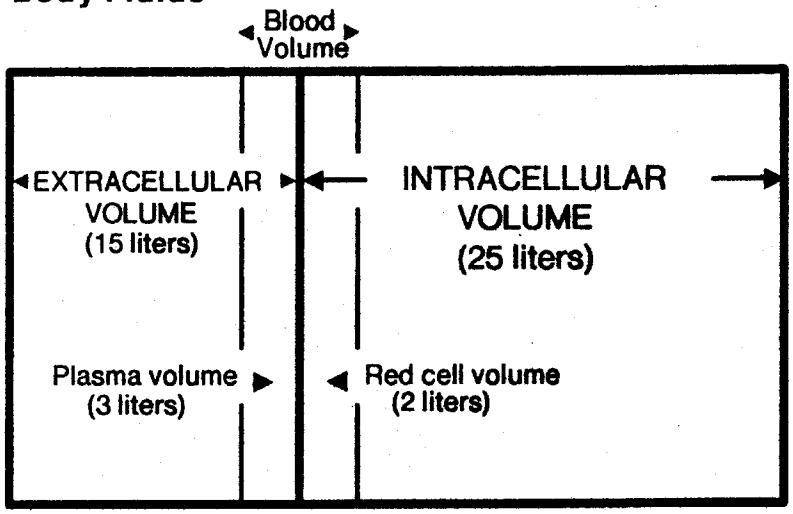
5.12 Collecting Tubule

The collecting tubule was referred to earlier as the site at which ADH causes water reabsorption. The corresponding explanation box is shown in figure 20.

5.13 Tissues

The fluid present in tissues is either inside or outside cells and is therefore termed *intracellular* or *extracellular*, respectively. Figure 19 shows how the body's fluids may be separated into its subcomponents. *Plasma volume* is that component of extracellular fluid within blood vessels. The remaining portion of the extracellular fluid volume is called *interstitial* fluid, that fluid which lies between cells. In the diagram, the component of *intracellular* fluid contained within red blood cells is shown.

Fig. 19. The Body Fluids



(Adapted from Guyton's Textbook of Medical Physiology[10])

A causal context for normal Tissues is depicted in Figure 21. Since NEPHROS assumes normal tissue hydrostatic and oncotic pressures, interstitial fluid volume is determined by capillary hydrostatic and oncotic pressures. The component of plasma volume which does not enter the interstitial fluid space is returned to the heart. The user is presented with the calculated value of interstitial fluid volume and is given the opportunity to directly set the value of venous blood return to the heart. Also represented in figure 21 is the conversion of renin to aldosterone, which was described above.

Fig. 20. The Collecting Tubule Explanation Box

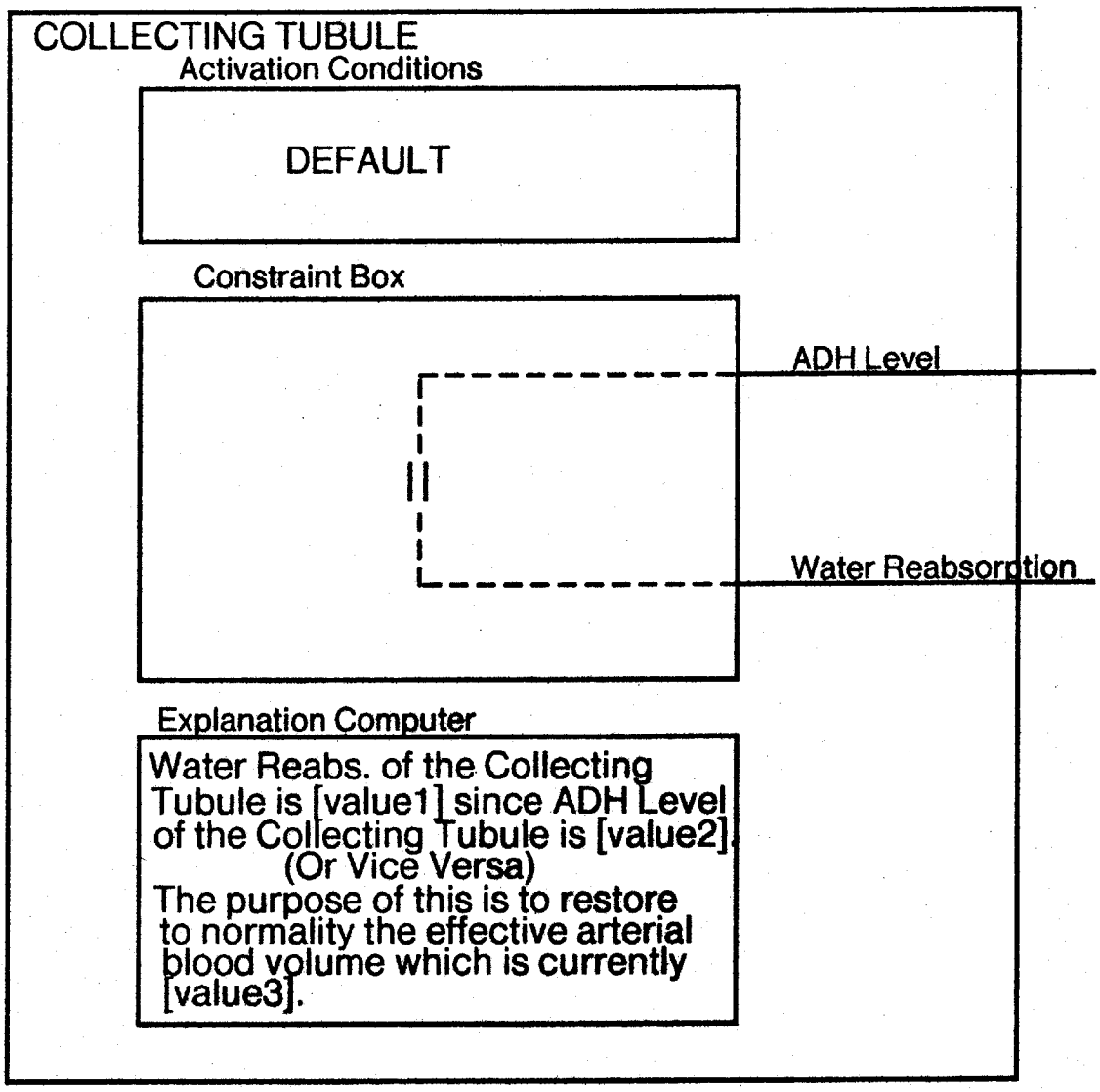
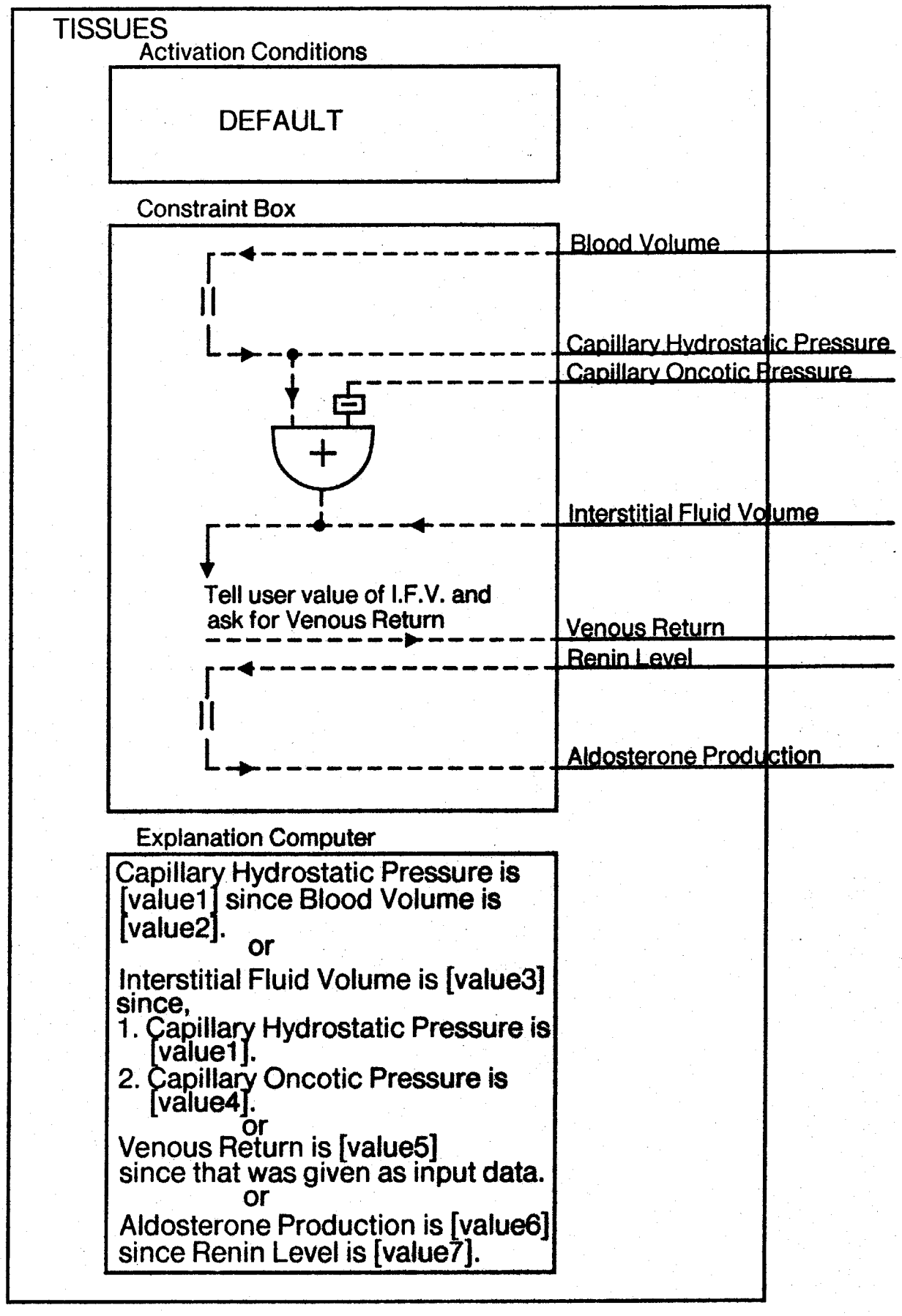


Fig. 21. The Tissues Explanation Box



6. Three Syndromes

In the last chapter, most of the explanation boxes used by NEPHROS to represent normal bodily function were shown. Here the pedagogical network shall be expanded to include the pathophysiology underlying hypoperfusion of the kidney, SIADH and nephrotic syndrome. We'll begin with a discussion of heart failure, one of the causes of kidney hypoperfusion.

6.1 Heart Failure

Heart failure is considered to be the "pathophysiological state in which an abnormality of cardiac function is responsible for failure of the heart to pump blood at a rate commensurate with the requirements of the metabolizing tissues"[2]. Here we shall assume that tissues are metabolizing at a normal rate and thus the term heart failure corresponds to decreased cardiac output.

We have a physiological version of Ohm's Law which tells us:

Blood Pressure = Cardiac Output x Total Peripheral Resistance

The decreased cardiac output seen in heart failure will therefore cause a decreased arterial blood pressure if total peripheral resistance is within normal limits. The juxtaglomerular apparatus of the kidney senses this decreased blood pressure and responds with increased renin production. By the rather circuitous route mentioned earlier, the tissues convert renin into aldosterone. The aldosterone then enters the circulatory loop and finds its way to the distal convoluted tubule of the kidney to cause sodium retention. Thirst and ADH secretion are stimulated since osmolarity of the blood is increased and arterial blood pressure is decreased. The increased ADH level stimulates water reabsorption in the collecting tubule of the kidney, while the thirst complex stimulates water intake. The retained fluid and the increased water intake cause an increased circulating blood volume, which is reflected by an increased capillary hydrostatic pressure in the tissues. This pressure imbalance causes edema by forcing fluid from the blood into the tissues.

This causal rationalization is suitably termed *forward heart failure*, the hypothesis set forth by MacKenzie in 1913 which describes the physiological manifestations of heart failure as result of deficient expulsion of blood into the arterial vessels. For this thesis, we will join the ranks of MacKenzie. It should be mentioned that in 1832, James Hope referred to the importance of back pressure generated in the circulatory system by poor blood ejection from the heart. This concept of *backward heart failure* describes how buildup of venous pressure can cause edema. It is likely that both mechanisms are operable in most patients with the signs and symptoms of chronic heart malfunction[13].

In the interaction with NEPHROS portrayed earlier, the physiology underlying heart failure provided an illustration of the program's performance. The normal HEART and the HEART IN FAILURE explanation boxes mentioned during the interaction are shown in figures 22 and 23, respectively. The EDEMATOUS TISSUES explanation box which was later activated is shown in figure 24.

Some researchers have described a second mechanism for sodium retention in heart failure which acts alongside the renin-angiotensin-aldosterone system. By physiological processes not completely understood, an increased retention of sodium by the proximal part of the nephron is often observed in situations of decreased effective arterial blood volume [2]. The explanation box of figure 25 is a representation of this phenomenon. By permitting this causal description in the pedagogical network, the causality relation C must be extended to include relationships of the form **a C b**, where event **b** is *associated* with event **a**.[1]

---

1. The reader may wonder what happened to the three input IQ-adder shown earlier which compactly described the factors influencing sodium retention. In heart failure, a decreased GFR is caused by the decreased effective arterial blood volume. This dependency is hidden within the relationship shown in figure 25. The effect of natriuretic hormone has not been included in NEPHROS due to current lack of knowledge of its properties.

Fig. 22. Normal Heart Explanation Box

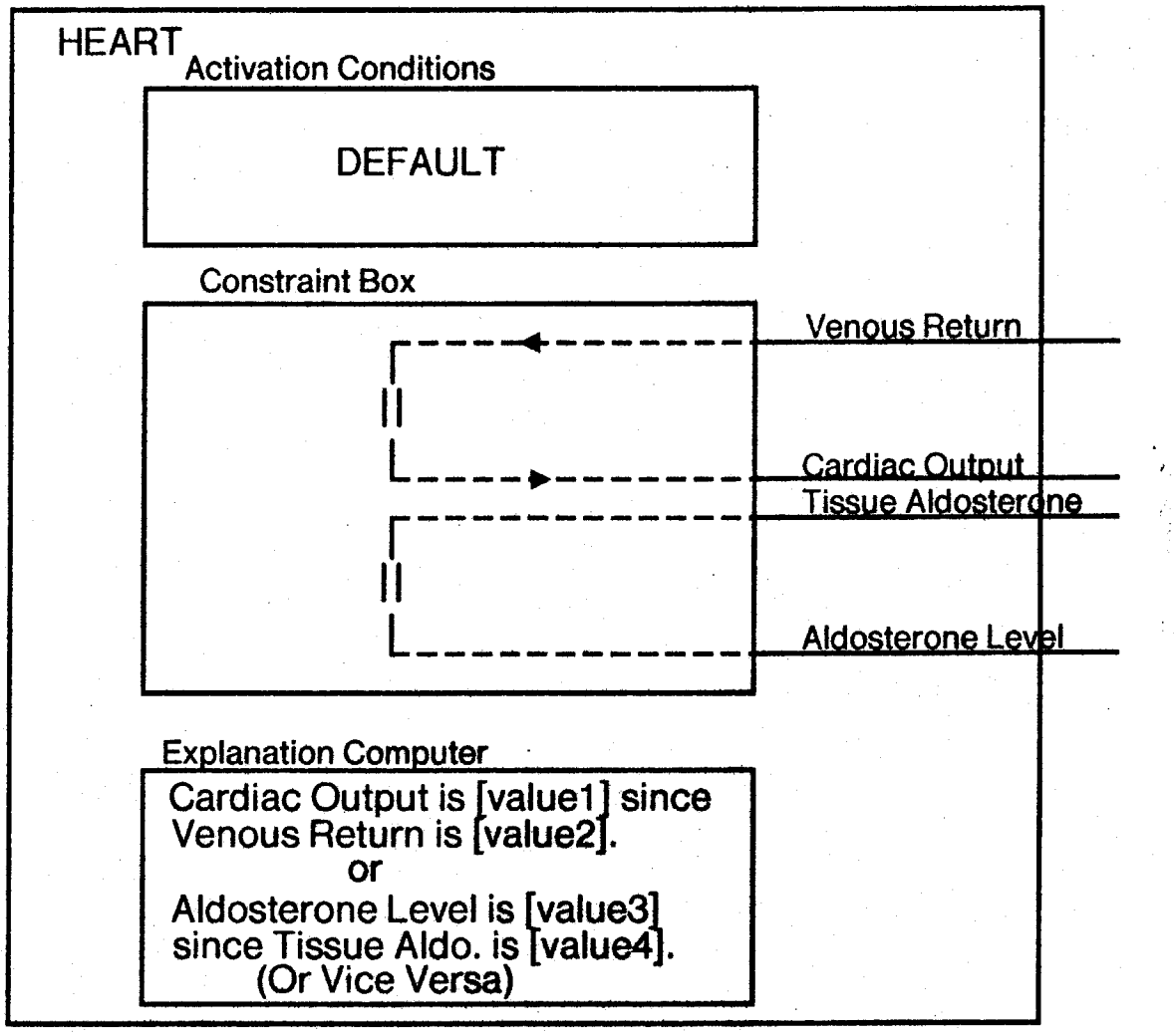


Fig. 23. Heart Failure Explanation Box

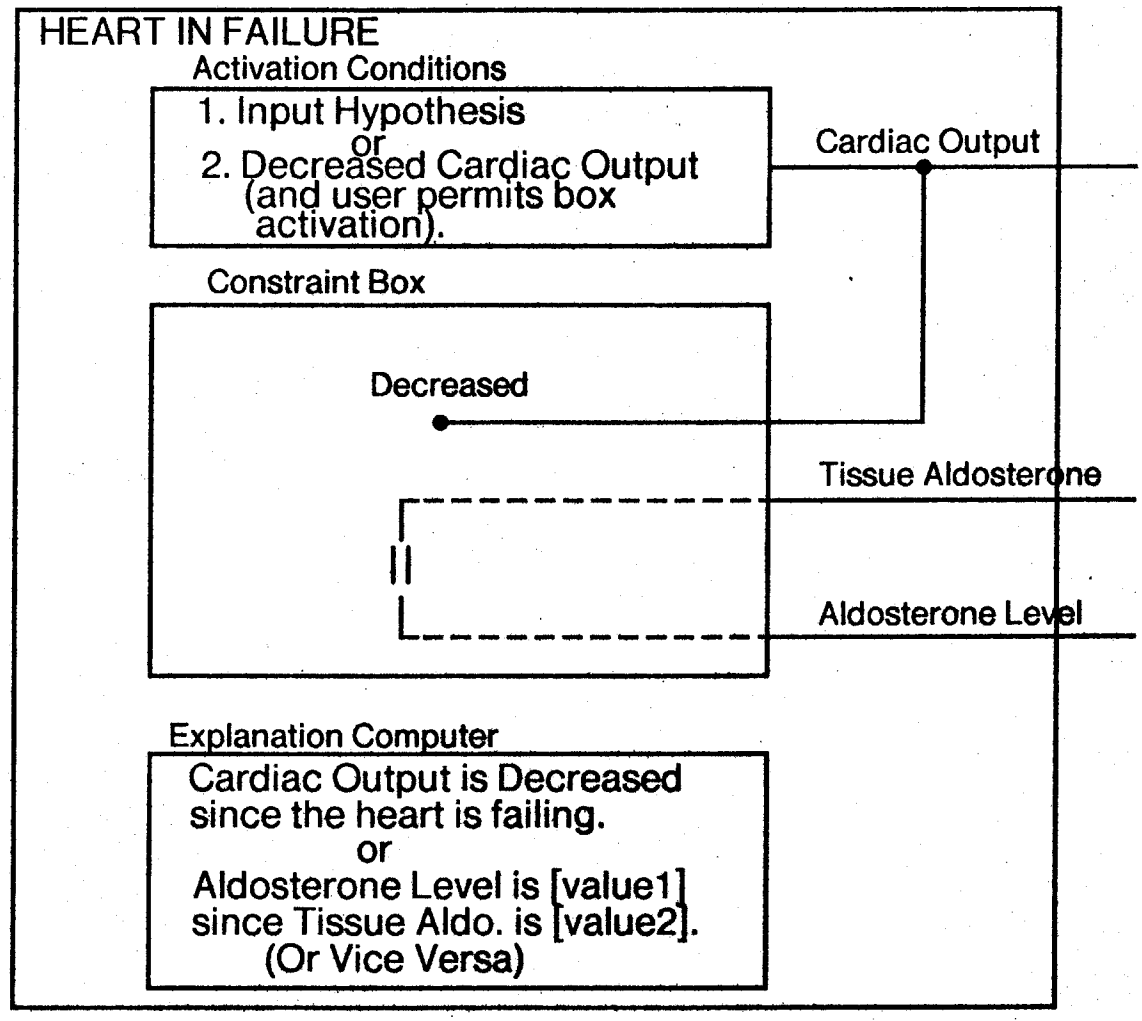


Fig. 24. Edematous Tissues Explanation Box

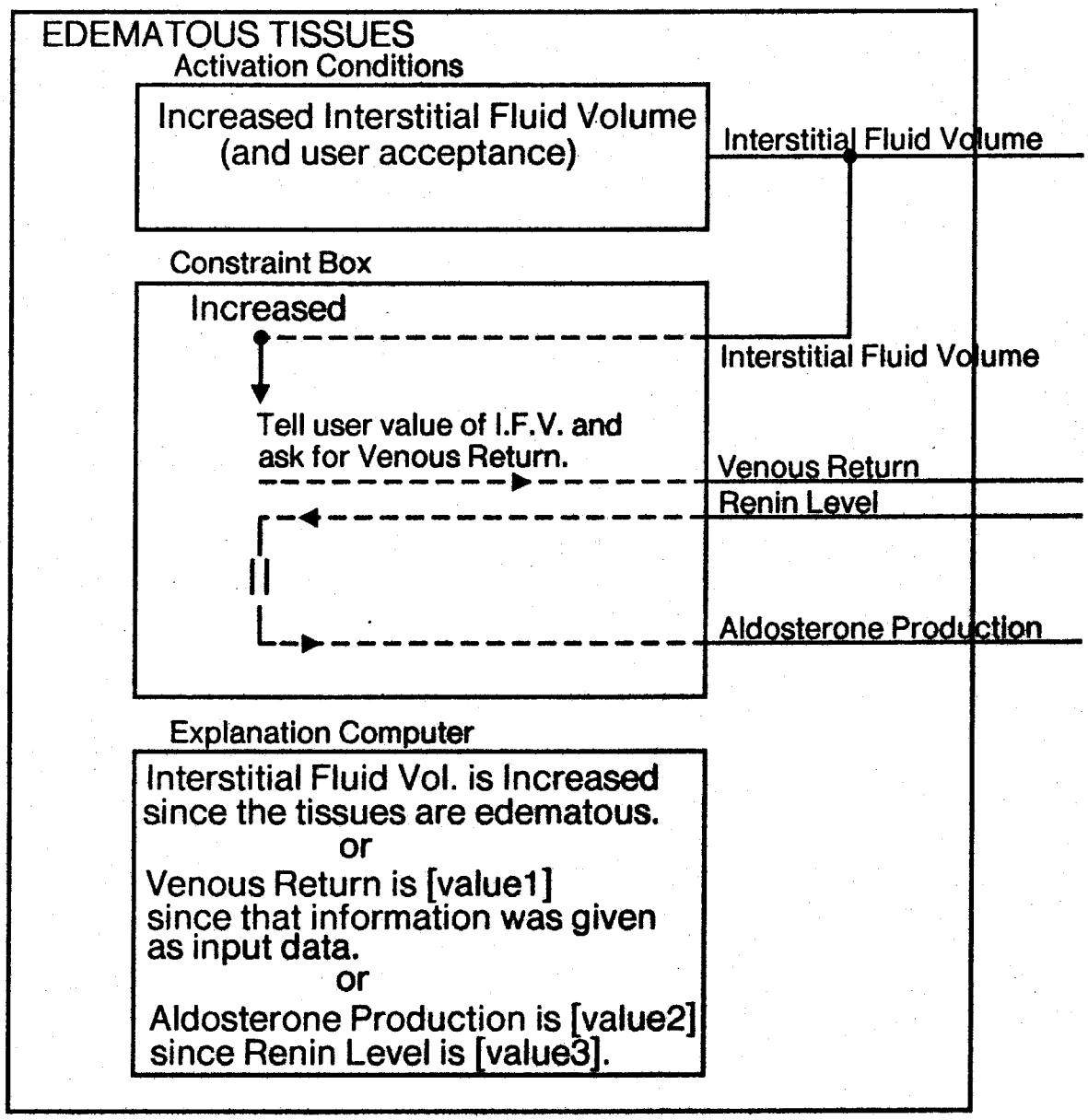
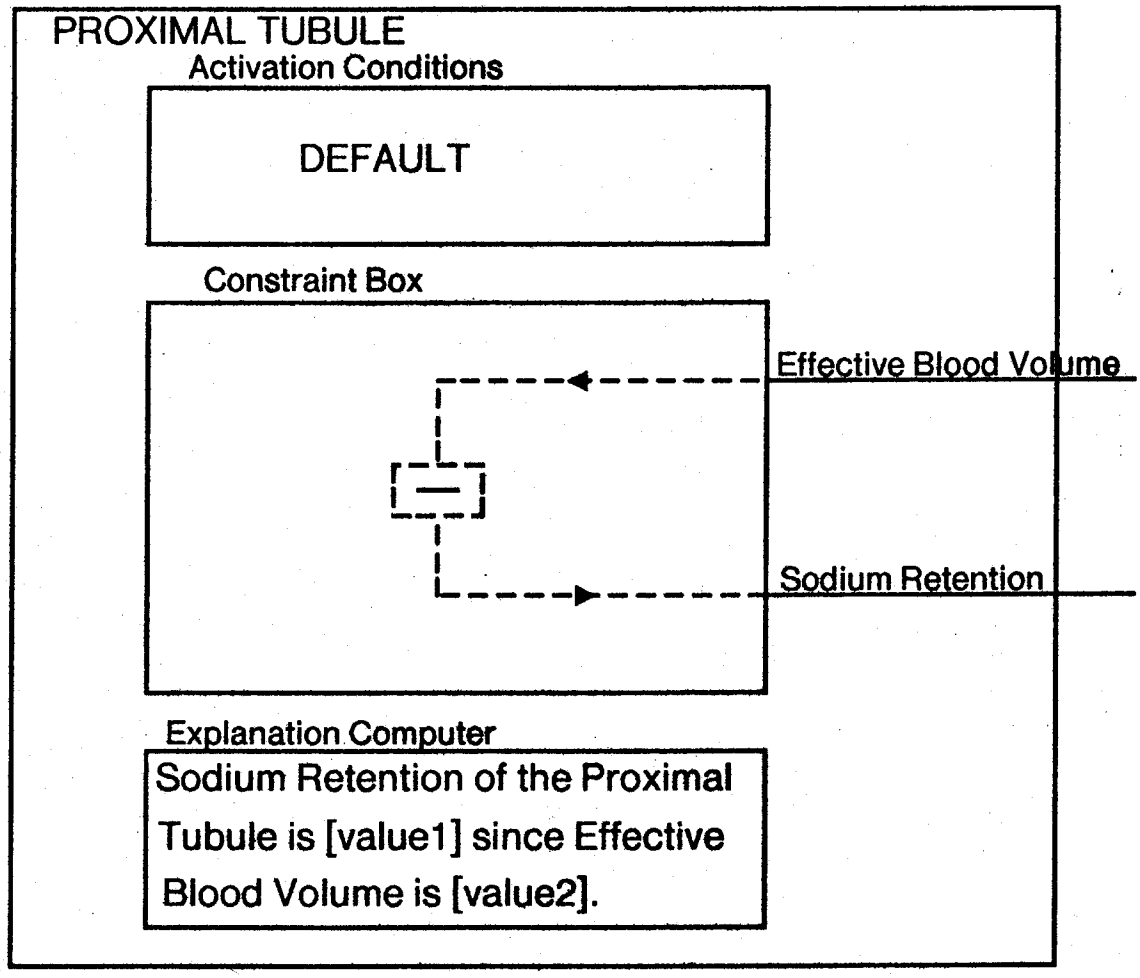


Fig. 25. The Proximal Tubule Explanation Box



6.2 SIADH

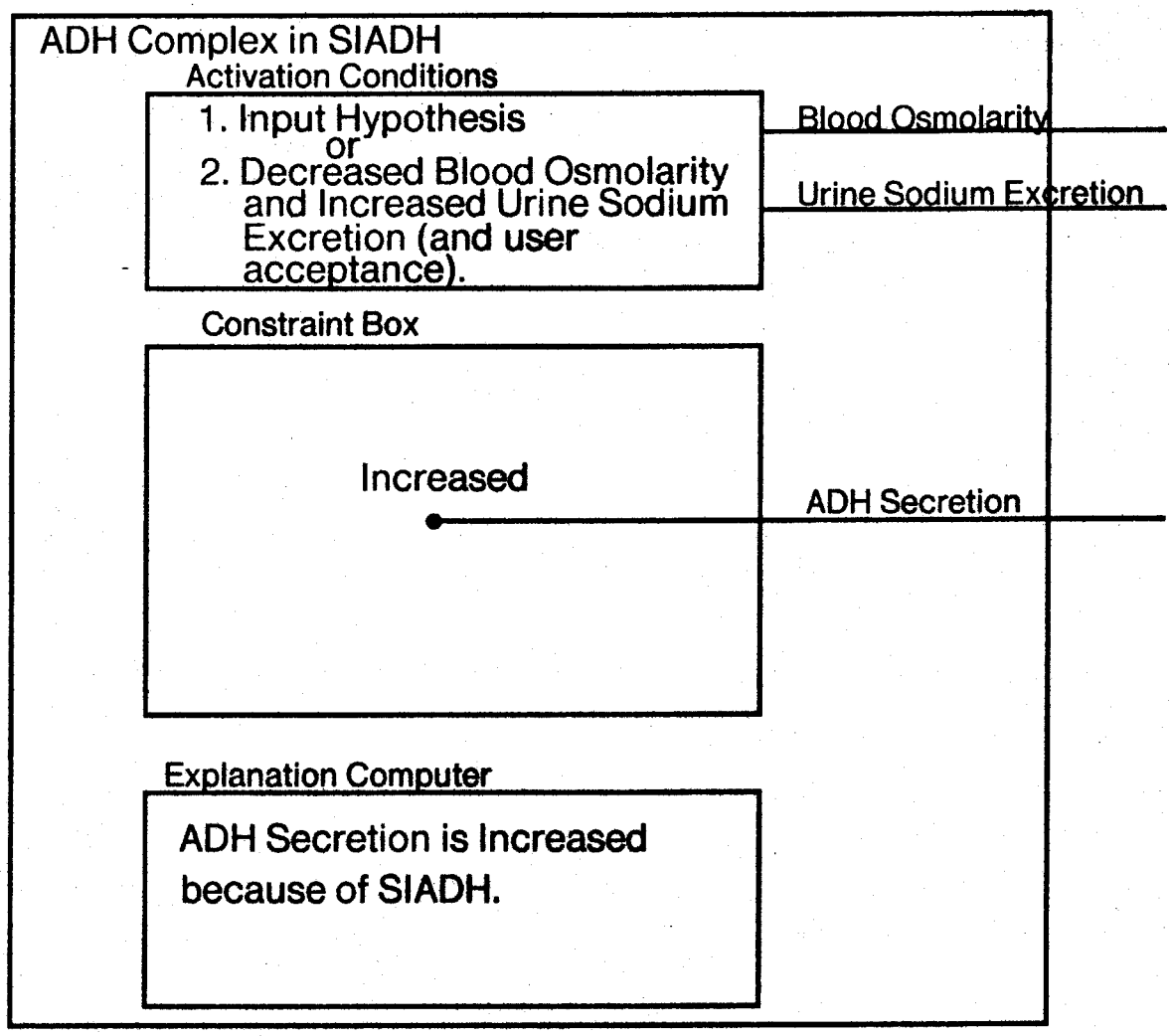
SIADH, the *s*yndrome of *i*nappropriate *s*ecretion of *a*ntidiuretic *h*ormone, is the condition in which ADH secretion is excessive in comparison to the requirements dictated by blood osmolarity and volume. The source of inappropriate secretion of ADH is usually cancerous lung. However, often it is the hypothalamus itself which is responsible for the hypersecretion. For sake of simplicity, we will view the etiology of SIADH as excessive production of hormone by the ADH complex. The explanation box of figure 26 describes the corresponding causal context. If we compare this diagram to that of the default ADH Complex explanation box shown earlier, it is apparent that arterial blood pressure and blood osmolarity are no longer influential in determining ADH secretion.

The consequence of the hormonal overproduction is water reabsorption in the collecting tubule above the body's needs. In modelling this process, NEPHROS will determine that an increased ADH level causes increased water retention and this information is passed to the TISSUES explanation box. The program then prompts us to give a value for venous return to the heart, which we logically set to *increased*. The resulting increased cardiac output, with the assumption of normal total peripheral resistance, causes an increased blood pressure. This increased pressure is sensed by the juxtaglomerular apparatus of the nephron, which responds by producing renin at a level below the norm.

The renin-angiotensin-aldosterone system will cause the distal tubule of the kidney to decrease its reabsorption of sodium. To NEPHROS, a decreased *proximal* tubular reabsorption of sodium is consistent with the increased effective arterial blood volume. The program determines that the net effect of decreased proximal and distal reabsorption of sodium is increased urine sodium excretion, as is usually the case in SIADH.

The activation conditions of the explanation box in figure 26 require mention. Since *decreased* blood osmolarity and *increased* urine sodium excretion are indicative of inappropriate ADH secretion, they are considered to be conditions which permit the ADH Complex in SIADH explanation box to be active. This causal context is automatically activated if SIADH is given as an hypothesis during the program's input phase.

Fig. 26. ADH Complex in SIADH Explanation Box



6.3 Nephrotic Syndrome

In the nephrotic syndrome, the kidney's glomerular basement membrane is malfunctioning. The disrupted filtration sieve permits the amount of protein lost in the urine to increase enormously. Since the level of protein in the blood decreases, blood oncotic pressure decreases. This imbalance of oncotic pressure causes fluid exudation into the interstitial fluid space. Since interstitial fluid is excessively increased, the tissues are considered edematous.

What is left of the fluid component of the blood is returned to the heart. The decreased venous return causes the output of the heart to decrease and the sequence of events described in the heart failure section above are retraced. The above pathophysiological description can be modelled by the NEPHROS hierarchy, with the addition of an ABNORMAL GLOMERULUS explanation box. The default GLOMERULUS and the ABNORMAL GLOMERULUS explanation boxes are shown in figures 27 and 28, respectively.

Fig. 27. Normal Glomerulus Explanation Box

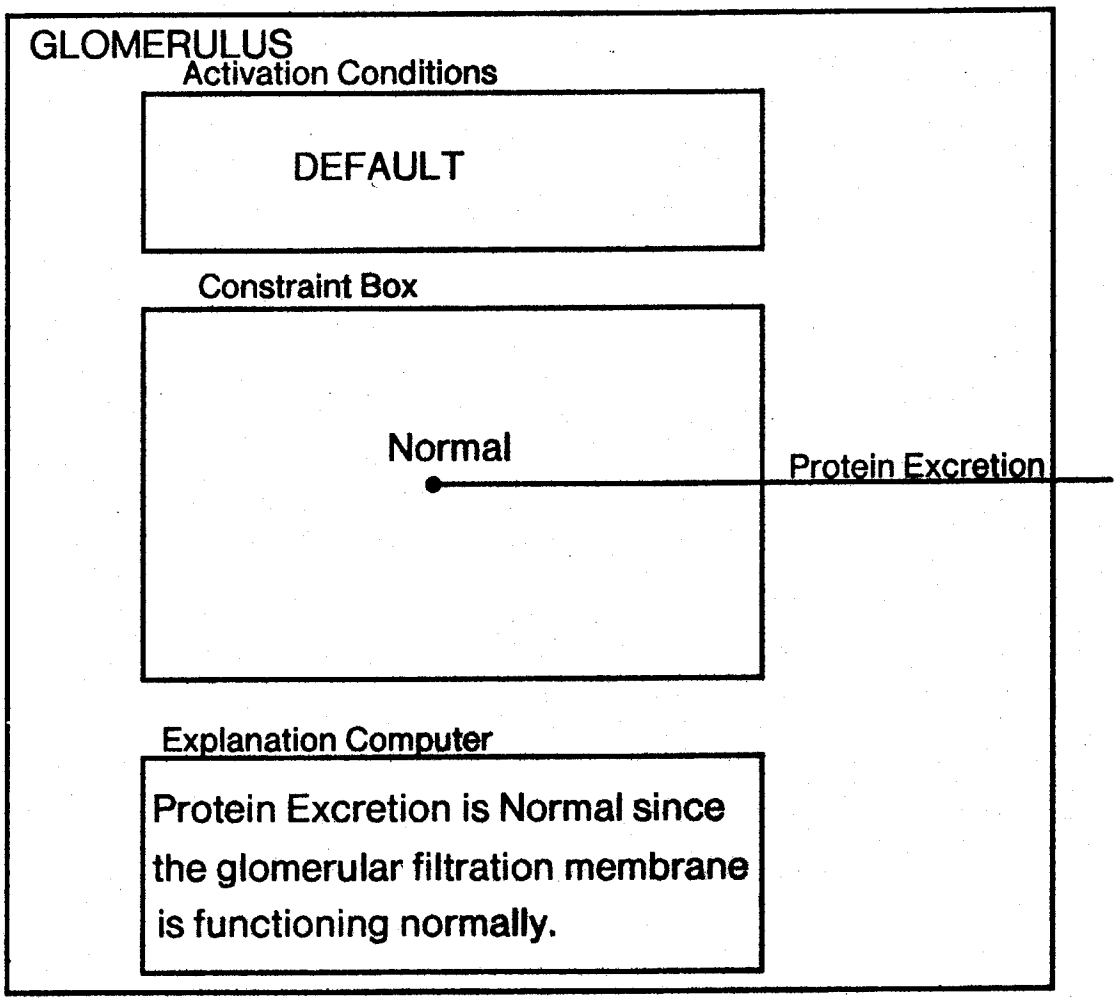
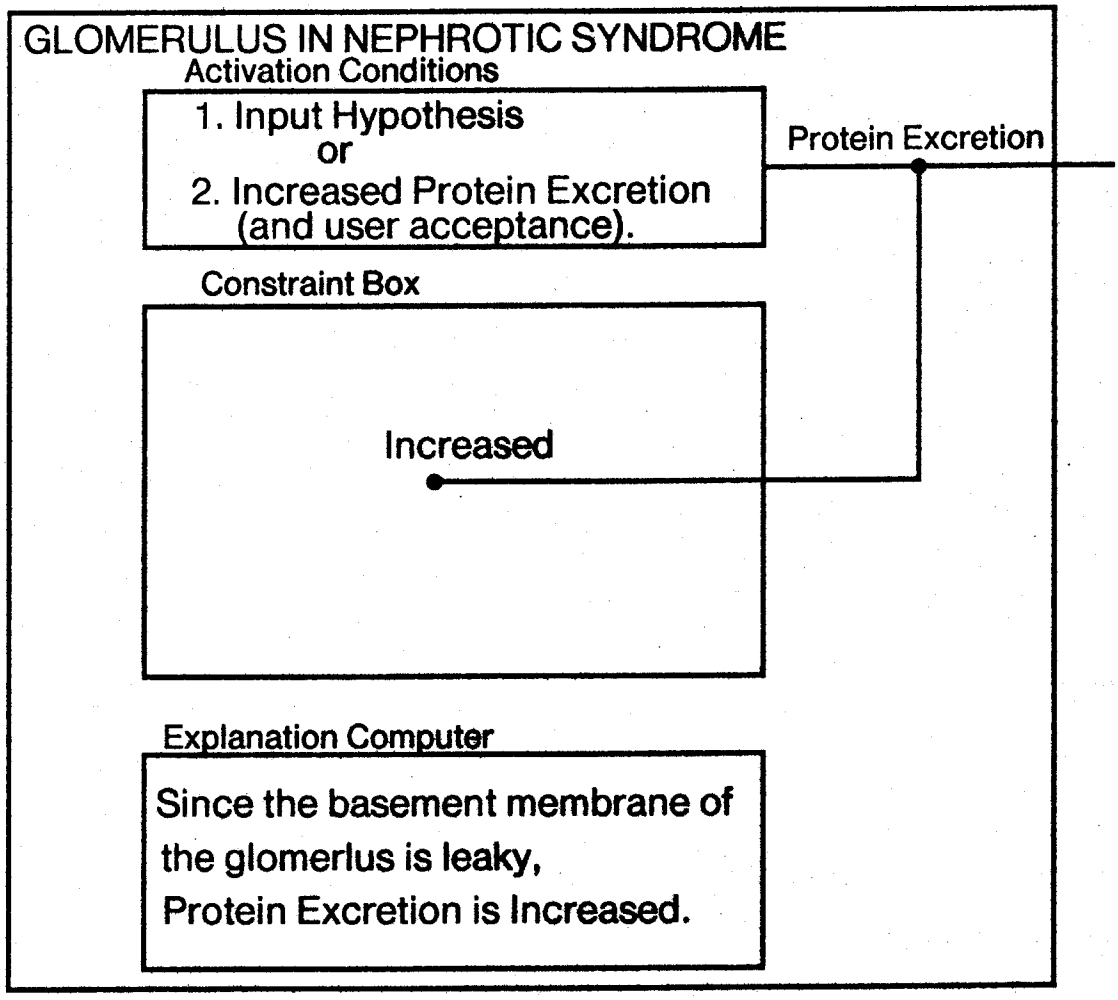


Fig. 28. Glomerulus In Nephrotic Syndrome



7. Epilogue

In this final chapter, some important points relevant to the theory of physiological modelling presented earlier will be mentioned. The report concludes with a list of possible directions for further research. The reader is encouraged to pursue any research issues described below.

7.1 Frames

When constructing the first draft of this thesis, I began to wonder whether the explanation box concept was in fact an instantiation of the *frame theory* introduced by Marvin Minsky. He describes the essence of his theory as follows:

> "...When one encounters a new situation (or makes a substantial change in one's view of the present problem) one selects from memory a substantial structure called a frame. This is a remembered framework to be adapted to fit reality by changing details as necessary.
>
> A frame is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party."[16]

We can think of the causal contexts described by explanation boxes as frames. As an example, the kidney explanation box models the framework we select from our memory when describing causality within the context of that organ. Other properties Minsky attributes to frames are of interest to us:

> "...We can think of a frame as a network of nodes and relations. The 'top levels' of a frame are fixed, and represent things that are always true about the supposed situation. The lower levels have many *terminals* – 'slots' that must be filled by specific instances or data.
>
> ...A frame's terminals are normally already filled with 'default' assignments.
>
> ...The frame systems are linked, in turn, by an information retrieval network. When a proposed frame cannot be made to fit reality – when we cannot find terminal assignments that suitably match its terminal marker conditions – this network provides a replacement frame."[16]

What is the explanation box analogue of terminals? Activation conditions, constraint box and explanation computer can be thought of as slots which must be filled in order to characterize a

particular causal environment. If the constraint box directly carries out constraint computations, those procedures we get by default. If pointers to lower level nodes in the explanation hierarchy are given, they can be thought of as slots which must be filled in with explanation boxes whose activation conditions are satisfied.

There may be terminal slots other than those mentioned above which are useful for causal context specification. The reader is encouraged to examine the article cited above to consider which additional information might be important.

7.2 The First Derivative

Increased, decreased, normal and *unknown* are the only values NEPHROS will assign to physiological variables. Is this IQ algebra powerful enough to describe all qualitative physiological envisionments? Minimally, we should expand our algebra to describe the rate of change of variables.

An envisionment of the consequences of heart failure was given in the last chapter. *Increased* water retention is a physiological response to the *decreased* effective arterial blood volume. The retained fluid causes capillary hydrostatic pressure in the tissues to increase and thus the tissues become edematous. Since the knowledge of increased fluid retention was passed directly to the tissues, we did not have to worry about how to add the fluid retained by the kidney to the effective blood volume. We should be able to represent the fact that effective blood volume is *decreased*, but it is currently *increasing*.

The problem of formalizing rate of change in the causal description of physical systems is being examined by Forbus[9] and Kuipers[14].

7.3 Quantitative Values

The main thrust of this thesis has been the presentation of a methodology for representing qualitative physiological relationships in a computer. However, this is but part of the solution to the development of complete physiological models. Although reasoning with qualitative information is very important, physicians are faced with a great deal of quantitative knowledge every day.

When analyzing a problem involving quantitative information, we may at some point proceed to analyze and discuss it using qualitative reasoning. Developing theories on how this internal conversion takes place is extremely important to the physiological modelling problem. This issue should provide many interesting research problems.

7.4 Reverse Constraint Networks

The causality relation C has been defined to give us a descriptive handle on constraint propagation. It was stated that in certain circumstances **a** $-_{causes}\rightarrow$ **b** may be in the direction opposite to physiological causality. Let us consider what would happen if we carried this method of deriving information to the extreme and built constraint networks whose main computational power was backwards propagation.

The current version of NEPHROS permits both clinical manifestations and hypotheses as input data and the main task of the program is to determine physiological consequences of the data. It would be interesting to investigate whether a similar program could perform envisionment in the reverse direction. If it is known that "effective arterial blood volume is decreased", which we will label as event **b**, the program should be able to determine all possible events **a** such that **a** causes **b** in the physiological sense. In such a program, generation of a list of physiological disorders which could have caused the patient's signs and symptoms would be the end result and therefore diagnosis would be the main focus.

7.5 General Directions For Research

Possible applications of the theory presented in this thesis and directions for further research are delineated below:

1. Refinement and expansion of NEPHROS to include acid base metabolism.

2. Building a constraint network which represents the kidney at a deeper level of causal description. Modelling Henle's loop, the countercurrent mechanism, ion exchange, peritubular hemodynamics and so forth should be a formidable task.

3. Development of constraint networks for other physiological systems, such as the respiratory system.

4. Construction of pedagogical networks for electronic circuits, computer networks and other hierarchical systems amenable to the grey box approach.

7.6 Conclusion

In this thesis, a theory for representing in the computer the physician's mental model of physiology has been introduced. Building formal descriptions of the mechanisms underlying human physiology will enable us to create medical computer programs with admirable diagnostic and therapeutic power. If the advice and corresponding justifications produced by such a program is acceptable, the physician – as the final integrator – could combine his or her expertise with that information. With the current constraints on our modern day health care system, this could provide substantial improvement in patient care.

Appendix I - Glossary Of Medical Terms

Unless indicated to the contrary, the following definitions are derived from those contained in Stedman's Medical Dictionary[3].

ADH *Antidiuretic hormone.*

Anatomy The science of the morphology or structure of organisms.

Arterial Relating to one or more arteries or to the entire system of arteries.

Arteries Blood vessels conveying blood away from the heart.

Capillary A tiny blood vessel which exchanges fluid, nutrients, electrolytes, hormones and other substances between the blood and the interstitial spaces[10].

Cardiac Pertaining to the heart.

Cardiac Output The rate of blood ejection from the heart.

Concentration The quantity of substance per unit volume of solution.

Distal Farthest from the trunk or point of origin. Opposite of proximal.

Edema An accumulation of an excessive amount of fluid in the tissues.

Effective Blood Volume The amount of blood that effectively supplies the tissues. A variation of *effective circulating volume*[19].

Electrolyte A substance which dissociates into ions when dissolved in water.

Glomerulus A tuft of small blood vessels at the beginning of each functional subunit (nephron) of the kidney.

Homeostasis The maintenance of static, or constant, conditions in the internal environment[10].

Hormonal Pertaining to hormones.

Hormone A chemical substance, formed in one organ or part of the body and carried in the blood to another organ or part.

Hypoperfusion Decreased blood flow to the tissues under consideration.

Oncotic Pressure The osmotic pressure generated by proteins in solution[10].

Osmolarity The number of osmoles of solute per liter of solution.

Osmole One of the particles generated by a substance when it is dissolved in a solution. One gram mole of nondiffusible and nonionizable substance is equal to 1 osmole[10].

Osmotic Pressure The tendency for fluid to flow between two solutions which are separated by a semipermeable membrane. The pressure difference is determined by differences in osmolarity of the solutions.

Patho- Pertaining to disease

Pathology The study of disease.

Prognosis A forecast of the outcome of a disease.

Proximal Nearest the trunk or point of origin.

Renal Relating to the kidneys.

Sign An abnormality discovered by the physician during a physical examination.

Symptom A phenomenon experienced by the patient which is indicative of disease.

Syndrome The aggregate of signs and symptoms associated with any morbid process.

Vein A blood vessel carrying blood toward the heart.

Venous Relating to a vein or to the veins.

Appendix II - A Theorem On Partial Orderings

Definition Let R be a relation. The relation R^{-1} (pronounced "R inverse") is defined by:

$xR^{-1}y$ if and only if yRx.

Theorem If R is a strict partial ordering on set X, R^{-1} is also a strict partial ordering on set X.

Proof of theorem[1] Let R be a strict partial ordering on X.

(a) *Irreflexivity of R^{-1}:*

Let x be an arbitrary member of set X.

Assume, per absurdum, $xR^{-1}x$.

By definition of R^{-1}, we have xRx.

Since R is irreflexive, this is disallowed.

Therefore, by indirect proof, $\forall x \in X [\neg(xR^{-1}x)]$.

(b) *Antisymmetry of R^{-1}:*

Let x and y be arbitrary members of set X.

Assume $xR^{-1}y$.

By definition of R^{-1}, we have yRx.

Since R is antisymmetric, we have $\neg(xRy)$.

By definition of R^{-1}, we have $\neg(yR^{-1}x)$.

Therefore, by conditional proof, $\forall x, y \in X [xR^{-1}y \Rightarrow \neg(yR^{-1}x)]$.

(c) *Transitivity of R^{-1}:*

Let x, y and z be arbitrary members of set X.

Assume $zR^{-1}y$ and $yR^{-1}x$.

By definition of R^{-1}, we have yRz and xRy, respectively.

Since R is transitive, we have xRz.

By definition of R^{-1}, this gives us $zR^{-1}x$.

---

1. In response to exercise 7.2 in [7].

Therefore, by conditional proof, $\forall x, y, z \in X [zR^{-1}y \text{ and } yR^{-1}x \Rightarrow zR^{-1}x]$.

By (a) (b) and (c) above, R^{-1} is a strict partial ordering on X. □

Appendix III - Arterial Blood Computations

The computations performed by the explanation boxes pointed to within the Arterial Blood explanation box are given below. The equations are in qualitative algebra:

Explanation Box	Constraint Box Computation
Blood Volume	Cardiac Output → Effective Blood Volume
Blood Pressure	Blood Pressure = Cardiac Output x Total Peripheral Resistance
Blood Sodium	Kidney Sodium Retention → Blood Sodium
Blood Osmolarity	Blood Osmolarity = Blood Sodium / Effective Blood Volume
Blood Oncotic Pressure	Blood Oncotic Pressure = Blood Protein / Effective Blood Volume
Blood Renin	Blood Renin = Kidney Renin Production
Blood ADH	ADH Level = ADH Production
Blood Aldosterone	Blood Aldosterone = Heart Aldosterone Level
Adjusted Blood Volume	Adjusted Blood Volume = Water Intake – Kidney Water Excretion If Effective Blood Volume is not set: Adjusted Blood Volume → Effective Blood Volume

References

1. Bartter, F.C; Schwartz, W.B. *The Syndrome of Inappropriate Secretion of Antidiuretic Hormone*, American Journal of Medicine, Volume 42, May 1967.

2. Braunwald, E. Heart Disease, W.B. Saunders Company, Philadelphia, 1980.

3. Cutler, A.G.(ed.). Stedman's Medical Dictionary, The Williams and Wilkins Company, Baltimore, 1972.

4. de Kleer, Johan. *The Origin and Resolution of Ambiguities in Causal Arguments*, Proceedings of IJCAI-79.

5. de Kleer, Johan. Qualitative and Quantitative Knowledge in Classical Mechanics, M.I.T. AI Technical Report 352, December 1975.

6. de Kleer, Johan. Causal and Teleological Reasoning in Circuit Recognition, M.I.T. AI Technical Report 529, September 1979.

7. Enderton, Herbert B. Elements of Set Theory, Academic Press, Inc., New York, 1977.

8. Forbus, Kenneth. *A CONLAN Primer*, Technical Note, November 1980, Bolt Beranek and Newman Inc., Cambridge, Mass.

9. Forbus, Kenneth. *Qualitative Process Theory*, M.I.T. AI-Memo 664, February 1982.

10. Guyton, Arthur C. Textbook of Medical Physiology, W.B. Saunders Company, Philadelphia, 1981.

11. Guyton, A; Jones, C; Coleman, T. Circulatory Physiology: Cardiac Output and its Regulation, W.B. Saunders Company, Philadelphia, 1973.

12. Hofstadter, Douglas R. GÖDEL, ESCHER, BACH: an Eternal Golden Braid, Vintage Books, New York, 1980.

13. Isselbacher, Adams, Braunwald, Petersdorf, Wilson. Harrison's Principles of Internal Medicine, Ninth Edition, McGraw-Hill Book Company, New York, 1980.

14. Kuipers, B. *Commonsense Reasoning About Causality: Deriving Behaviour From Structure*, Tufts University Working Papers in Cognitive Science, Number 18, May 1982.

15. Minsky, M; Papert, S. Perceptrons, The MIT Press, Cambridge, Massachusetts, 1969.

16. Minsky, Marvin. *A Framework for Representing Knowledge*, in The Psychology of Computer Vision by Patrick Winston, McGraw-Hill Book Company, New York, 1975.

17. Nilsson, Nils J. Principles of Artificial Intelligence, Tioga Publishing Co., Palo Alto, California, 1980.

18. Patil, Ramesh. Causal Representation of Patient Illness for Electrolyte and Acid-Base Diagnosis, M.I.T. LCS Technical Report 267, October 1981.

19. Rose, B.D. Clinical Physiology of Acid-Base and Electrolyte Disorders, McGraw-Hill Book Company, New York, 1977.

20. Shortliffe, E.H. Computer-Based Medical Consultations: MYCIN, American Elsevier Publishing Company Inc., New York, 1976.

21. Steele, Guy L. The Definition and Implementation of a Computer Programming Language Based on Constraints, M.I.T. AI Technical Report 595, August 1980.

22. Sussman, Gerald J; Stallman, Richard M. *Heuristic Techniques in Computer-Aided Circuit Analysis*, IEEE Transactions of Circuits and Systems, vol. CAS-22 (11), November 1975.

23. Sussman, Gerald J; Stallman, Richard M. *Forward Reasoning and Dependency-Directed Backtracking in a System for Computer-Aided Circuit Analysis*. Artificial Intelligence 9 (1977), 135-196.

24. Sussman, Gerald J; Steele, Guy L. *Constraints*, M.I.T. AI-Memo 502, November 1978.

25. Sussman, G.J. *Slices at the Boundary Between Analysis and Synthesis*, in Artificial Intelligence and Pattern Recognition in Computer Aided Design, edited by Latombe, North-Holland Publishing Company, 1978.

26. Szolovits, Peter(ed). Artificial Intelligence in Medicine, *AAAS Selected Symposium Series*, Westview Press, 1982.

27. van Dalen, D.; Doets, H.C.; de Swart, H. Sets: Naïve, Axiomatic and Applied. Pergammon Press Ltd., 1978.

28. Waltz, David. *Understanding Line Drawings of Scenes with Shadows*, in The Psychology of Computer Vision by Patrick Winston, McGraw-Hill Book Company, New York, 1975.