FOUNDATIONS OF ENVISIONING[1]

Johan de Kleer and John Seely Brown
XEROX PARC
Cognitive and Instructional Sciences
3333 Coyote Hill Road
Palo Alto, California 94304
August, 1982

ABSTRACT

This paper explores a particular kind of qualitative reasoning, called "envisioning," that is capable of producing causal explanations for device behavior. It has been implemented in a computer program, ENVISION, which can analyze a wide variety of thermal, fluid, electrical, translational and rotational devices. Rather than present the technical details of the envisioning process, this paper examines the theoretical foundations upon which it is built. Many of these considerations are ones that any builder of qualitative reasoning systems must face. In particular, two vital considerations are *explanation* and *robustness*: What notion of causality can adequately explain device behavior? How can we have any confidence in the theory's analysis of a novel device?

INTRODUCTION

The theory of envisioning [1] [2] has two central characteristics. First, it is a physics. As such envisioning is not concerned with *post hoc* rationalization of observed behavior, but rather with constructing correct predictions of a device's qualitative behavior. Second, it is a theory of causality in that it can be used to produce complete causal explanations of how a device functions. Roughly speaking, explanations are complete (with respect to a given grain size of analysis) when they can be used to predict behavior of a device under novel conditions. Typical novel conditions include those in which the device suffers a casualty or is operated in a new context. Such a theory of causal, qualitative reasoning is important for both cognitive science and artificial intelligence.

Envisioning is a form of reasoning that produces a causal explanation for the behavior of a physical device by explaining how disturbances in the device propagate. Envisioning is often confused with qualitative simulation; the latter is only the simplest form of envisioning. In more complex cases, envisioning is primarily concerned with introducing and manipulating assumptions about device functioning while maintaining a notion of causality. Thus, envisioning is best thought of as a problem-solving method that uses the device's

---

[1] A revision of the version appearing in the Proceedings of the American Association for Artificial Intelligence; 1982 August 16-20; Pittsburgh.

structure to guide and control the propagation of "tokens" through the device in order to discover ways to integrate assumptions about states, state transitions, and causal interactions into a coherent picture of how the device functions.

A typical kind of physical mechanism we might envision is a pressure regulator (see Figure 1) whose purpose is to maintain a specific pressure even though line loads and pressure sources vary.

The ENVISION program analyzes the pressure regulator and produces this causal explanation (in order to save space we have compressed the explanation and stated it in English rather than the exhaustive and lengthy output format of ENVISION): *"An increase in source (A) pressure increases the pressure drop across the valve (B). Since the flow through the valve is proportional to the pressure across it, the flow through the valve also increases. This increased flow will increase the pressure at the load (C). However, this increased pressure is sensed by (D) causing the diaphragm (E) to move downward against the spring pressure. The diaphragm is mechanically connected to the valve, so the downward movement of the diaphragm will tend to close the valve, thereby pinching off the valve. Because the flow is now restricted the output pressure will rise much less than it otherwise would have and thus remain approximately constant."*

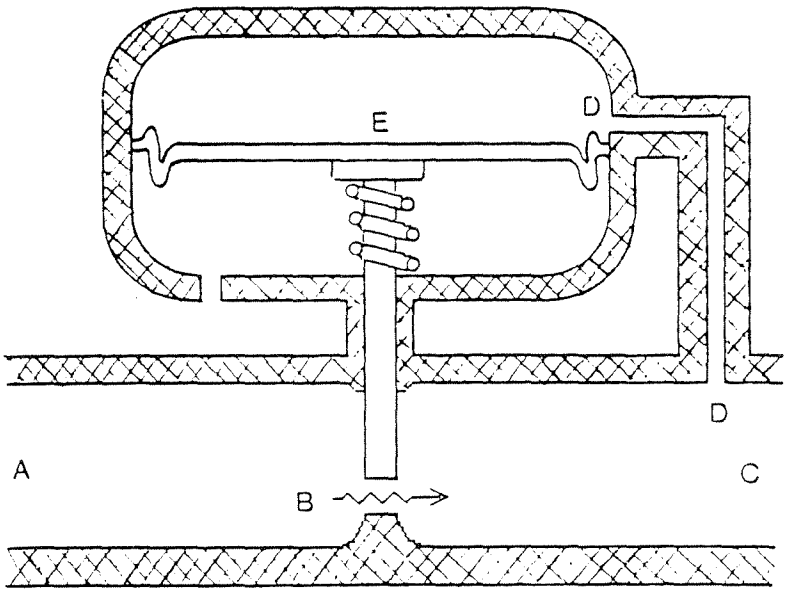

Figure 1 : Pressure Regulator

Although understanding the above explanation might seem simple, the analysis that the envisioning must perform to be able to support such explanations can be surprisingly complex. At a minimur envisioning must be able to determine the causal inputs and outputs for each component. This car an arduous task, especially when a component is connected by more than two ports to other comp

of the device. It must also detect and correctly determine the consequences of each kind of positive and negative feedback. Furthermore, since qualitative descriptions provide only partial information, it must be able to analyze underdetermined or underconstrained situations. Detailing the different kinds of reasoning strategies that enable the envisioning process to achieve these goals is not the subject of this paper (see [3] for a complete description of the problem-solving methods underlying envisioning). Instead, we examine the nature of the input evidence that the envisioning process operates on, the conclusions it produces, and the relationship between the two.

STRUCTURE AND FUNCTION (BEHAVIOR)

One objective of our investigation is to explore a theory of causal reasoning that can, given a physical device (in particular, a novel device), correctly predict ensuing behavior in that device. The devices are described by us, the investigators. This raises an enormous problem: even if the conclusions of the causal reasoning are correct, is its success attributable to the theory or to the way we encoded the device? (We assume that the envisioning system has available a library of abstract descriptions of the *behaviors* of device parts.) Certainly, the causal reasoning process will make deductions not present in the description of the situation, but the question remains whether these deductions form a *significant* portion of the total effort required to describe and analyze a physical situation. Is causal reasoning doing something valuable, or is most of the work it appears to be doing actually pre-encoded in the evidence provided to it?

One strategy to help ensure that the "conclusions" have not been pre-encoded into the evidence that the envisioning systems uses as input is to make the input evidence a well-defined ontological type, distinct from that of the conclusions. In particular, we require that the evidence be a description of the physical structure of the device, namely, its constituent parts and how they are attached to each other. The conclusions then describe the behavior, or functioning of the overall device. The task of causal reasoning (and of the envisioning system performing the reasoning) is to deduce the functioning of the device from its structure.

Part of the evidence is represented by the *device topology* (see Figure 2), in which nodes represent important components of the device and edges represent connections between them. Another part of the evidence is the general component model library: Each type of component and connection has a specific model which describes its behavior in the abstract, independent of any particular context. A component model describes all potential behaviors of the component in terms of qualitative equations on variables. For example, some important variables of a valve are the area available for flow, the pressure across the valve and the flow through it. By modeling each component, the abstract qualitative behavior of the overall device is implicitly characterized by a set of qualitative equations. An important qualitative equation for a valve is that the pressure across it is proportional to the flow through it (as described by the following section on component models). This set of equations is then "solved," and the solution interpreted in terms of the structure of the device. This solution process must be of a special kind so that causal explanations for its conclusions can be extracted from the solution process.

A second strategy to ensure that the conclusions are not pre-encoded in the evidence is to design the component models and the reasoning process to be "context-free." By being context-free, one library of models and one causal reasoning process should successfully analyze a wide variety of physical devices, particularly devices that have not yet been analyzed or devices operating under new conditions. We call this meta-theoretic constraint the *no-function-in-structure* principle. The class of devices constructible from any particular set of components is, in principle, infinite in number. We can only check our envisioning process on a finite subset of this infinite class; the no-function-in-structure principle provides some confidence that it will also succeed on the untested devices. The principle thus improves the descriptive adequacy of our causal theory.

Take as a simple example a light switch. The model of a switch that states, "if the switch is off, no current flows; and if the switch is on, current flows," violates the no-function-in-structure principle. Although this model correctly describes the behavior of the switches in our offices, it is false in general as there are many closed switches through which current does not necessarily flow (such as, for example, two switches in series). Current flows in the switch only if it is closed and there is a potential for current flow. One of the reasons why it is surprisingly difficult to create "context-free" descriptions of a component is that whenever one thinks of how a component behaves, one must, almost by definition, think of it in some type of supporting context. Thus, the properties of how the component functions in that particular supporting context are apt to subtly influence how one models it.

DEVICE TOPOLOGY

A device consists of constituents. Some of these constituents are components that themselves can be viewed as smaller devices (e.g., resistors, valves, boilers). Other constituents are connections (e.g., pipes, wires, cables) through which the components communicate by transmitting information. These connections can be thought of as conduits through which "stuff" flows, its flow described by conduit laws.

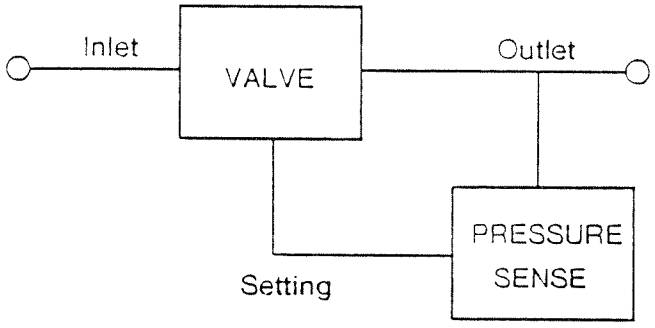

Figure 2 : Device Topology of the Pressure Regulator

Different types of conduits communicate different types of information. For example, the model for the pipe between the boiler and the turbine of a steam plant communicates pressure and steam, whereas

the model of a wire between a flashlight's battery and its light bulb communicates voltage and current. Most conduit types can be modeled by two attributes, one pressure-like and the other flow-like. For a fluid device, the two attributes are volumetric flow and pressure; for thermal devices, heat-flow rate and temperature; for translational devices, force and velocity; for rotational devices, torque and angular velocity; for electrical devices, current and voltage.

COMPONENT MODELS

Envisioning, in part, analyzes behavior in terms of qualitative disturbances from equilibrium. This motivates the class of values (used to denote increments of velocities, voltages, flows, etc.) to be positive, zero, and negative in order to represent the direction of the change, if any, from equilibrium. The value of every attribute must be encoded as one of "+," "0," "−" or "?"; no other choices are possible. Arithmetic with these values is straightforward. For example, if $x =$ "+" and $y =$ "+" then $x + y =$ "+" because if $x > 0$ and $y > 0$ then $x + y > 0$.

A component model characterizes all the potential behaviors that the component can manifest. It does not, however, specify which conduits connected to them are causal, that is inputs, and which are outputs; that can only be determined in the broader context of how a particular component is used in the overall device. The qualitative behavior of a valve (a component of the pressure regulator) is expressed by the following qualitative equation, called a "confluence": $AdP + PdA - dQ = 0$. In this equation Q is the flow through the valve, P is the pressure across the valve, A is the area available for flow, and dQ, dA and dP represent disturbances from the equilibrium values of Q, A and P. Given the situation in which the pressure across the valve is positive and area, as always is, positive, the expression simplifies (using the qualitative calculus just sketched out) to: $dP + dA - dQ = 0$. The confluence represents multiple competing tendencies, each encoding a different potential causal relationship. One such relationship is: if the area increases, but the flow remains constant, the pressure decreases.

A single confluence often cannot characterize the behavior of a component over its entire operating range. Thus, this range must be divided into subregions, each characterized by a different component state in which different confluences apply. For example, the behavior of the valve when it is completely open is quite different from that when it is completely closed. The notion of state is not strictly necessary in quantitative analysis since a single mathematical equation can adequately model the behavior of the component. Nevertheless it is often convenient to introduce state into quantitative analysis in order to delineate regions where certain effects are negligible or to form piece-wise approximations. On the other hand, in the qualitative regime the notion of state is absolutely necessary since it is not possible to formulate a *single* qualitative equation (i.e., confluence) which adequately characterizes the behavior of the component over its entire operating range.

The behavior of each qualitative state is provided by three types of rules. First, the component model specifies the region of operation covered by the component state. From these rules envisioning can

determine what transitions between states are plausible. For example, the closed state of a valve is defined by the condition $[A = 0]$, stating that if the component is in state closed there is no area available for flow and if there is no area available for flow the component is in state closed. Second, the component model provides confluences relating a component's variables. These rules are used to determine the states a component might be in and to test whether a transition *can* occur. Finally, the component model includes confluence equations that constrain component variables. Confluences describe the incremental behavior of the component and are used in constructing causal explanations for a device's behavior once the states of operation have been determined.

The full model for the valve is (as specified to ENVISION):

OPEN: $[A = A_{MAX}], P = 0, dP = 0$

WORKING: $[0 < A < A_{MAX}], P - Q = 0, dP + dA - dQ = 0$

CLOSED: $[A = 0], Q = 0, dQ = 0$;

From the state specifications, it is straightforward (and done automatically by ENVISION) to identify the possible state transitions:

OPEN: $dA = - \Rightarrow$ WORKING

WORKING: $dA = - \Rightarrow$ CLOSED, $dA = + \Rightarrow$ OPEN

CLOSED: $dA = + \Rightarrow$ WORKING.

In our earlier papers we used a formalism for component models that made these implicit facts explicit.

THE ENVISIONING PROCESS

Envisioning must determine (1) in which state(s) the overall device could be, (2) the causal behavior of the device in each of these consistent states, and (3) the possible transitions between pairs of global device states. Each kind of result has a different form of underlying reasoning and associated explanation. In the remainder of this paper we concentrate only on the second function of envisioning, namely the causal behavior, and on its explanation.

By modeling the behavior of each of the device's constituents, the potential behavior of the device is expressed as a set of confluences relating changes in variables (e.g., $dP + dA - dQ = 0$). The type of the device (thermal, electrical, etc.) and the types of the variables (velocity, current, etc.) become irrelevant. There are numerous techniques for finding solutions (assignments of values to variables) to the confluences of which relaxation, e.g., constraint satisfaction, is one. Although these techniques can correctly predict the behavior and satisfy the no-function-in-structure principle, most are incapable of yielding reasonable explanations for their predictions. For example, the best kind of explanation a constraint satisfaction technique can give for a solution is that the solution is an assignment of values consistent with the confluences of the component models. One needs to further analyze these confluences to construct a causal explanation for the behavior of the device.

The explanations produced by envisioning are based on a very simplistic notion of causality which we call *naive mechanism*. A naive mechanism's causal explanation consists of a series of effects on components, each of which is caused by previous effects on its neighboring components: E_1 (the initial disturbance) causes E_2 causes ... E_n. An effect always occurs as a consequence of, and therefore after, a cause. The consequences of an effect cannot immediately affect its causes. Causality concerns change and does not explain why the components are behaving the way they are, but rather how changes in these behaviors happen (i.e., how disturbances from equilibrium propagate). However, this view is overly simplistic as causal explanations can be used to explain how a quiescent state is reached, why the device stays in a state, and how a device in a quiescent state responds to disturbances.

Meeting the requirements of causality is facilitated by limiting envisioning to only using information locally available at the site of each component in the overall device where locality is defined with respect to the device topology. Of course, for most devices this restriction forces envisioning to tentatively hypothesize about global aspects of the local analysis, aspects that can be clarified at the end of the analysis. These tentative hypotheses are one of the major uses of assumptions. The next section describes the more commonplace use of assumptions, that of involving global ambiguity. In summary, constructing a qualitative reasoner that only makes predictions is relatively easy — the two difficult tasks are those of devising a consistent qualitative epistemology and constructing a reasoner which produces its predictions solely from its causal analysis.

ASSUMPTIONS AND PREDICTIVENESS

Because the information available to envisioning is qualitative, the actual behavior of the overall device may be underdetermined making more than one coherent behavior possible. Thus the concept of a correct behavioral prediction, which is central to our theory, needs to be spelled out. In order to analyze underdetermined situations, envisioning introduces explicit assumptions which it subsequently reasons upon. Thus, in underdetermined situations envisioning produces multiple interpretations, each with different assumptions and corresponding to a different overall device behavior.

Envisioning uses two fundamental modes of reasoning upon all of the assumptions. The first mode is straightforward. When a contradiction occurs during envisioning (e.g., different values assigned to the same variable), the underlying assumption set is marked as inconsistent and no interpretation may contain it. The introduction of an assumption introduces significant complexity to both the envisioning process and the explanations it produces. The second mode comprises strategies to try to minimize the number of assumptions necessary to find all the possible behaviors.

At a minimum, for a prediction to be correct, one of the interpretations must correspond to the actual behavior of the real device. A stronger criterion follows from observing that a structural description, abstracted qualitatively, of a particular device implicitly characterizes a wide class of different physically realizable devices with the same device topology. The stronger criterion requires that for the predictions

produced by envisioning to be correct then (1) the behavior of each device in the class is described by one of the interpretations, and (2) every interpretation describes the behavior of some device of the class.

This underdeterminacy has three immediate consequences. First, envisioning must be able to deal with underdetermined situations, a topic that in itself is difficult. Second, other external knowledge, perhaps of the teleology or known functioning of the actual device, is required to identify the correct interpretation for that particular device. Third, the notion of naive mechanism must be extended to include local assumption steps in causal explanations. (For a more detailed discussion of these points and their psychological ramifications see the latter sections in [1].)

CONCLUSION

Causal, qualitative reasoning is a difficult task; and ENVISION is a substantial program capable of producing valuable analyses for device behaviors that surprise even its implementors. A reasoning system should not be evaluated on the nature of its conclusions, but rather on the complexity of the relationship it establishes between its input and output. The input evidence to envisioning is structure — a device topology and the general library of component models. The output is function — device behavior and causal explanation of that behavior. The quality of the output is established by the predictiveness of the behavior and acceptability of the causal explanation. The no-function-in-structure principle merely ensures that structure and function are truly kept distinct.

ACKNOWLEDGMENTS

We thank Richard Fikes, Doug Lenat, Bob Lindsay, Tom Moran, Brian Smith, and Kurt VanLehn for their insightful discussions and comments.

BIBLIOGRAPHY

[1] de Kleer, J. and J.S. Brown, "Assumptions and Ambiguities in Mechanistic Mental Models," to appear in *Mental Models*, edited by D. Gentner and A. S. Stevens, Erlbaum, 1982.

[2] de Kleer, J. and J.S. Brown, "Mental Models of Physical Mechanisms and Their Acquisition," in *Cognitive Skills and their Acquisition*, edited by J.R. Anderson, Erlbaum, 1981.

[3] de Kleer, J. and J.S. Brown, "The Theory and Mechanics of Envisioning," forthcoming, Cognitive and Instructional Sciences, Xerox PARC, 1982.

[4] Forbus, K.D., "Qualitative Process Theory," Artificial Intelligence Laboratory, AIM-664, Cambridge: M.I.T., 1982.

[5] Forbus, K.D. and A. Stevens, "Using Qualitative Simulation to Generate Explanations," Report No. 4490, Bolt Beranek and Newman Inc., 1981.