FOUNDATIONS OF ENVISIONING

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

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 pay attention to. Two such considerations are explanation and robustness: What notion of causality is adequate for causal explanations of device behavior? How can there be any confidence in the analysis of a novel device?

INTRODUCTION

The theory of envisioning [1] [?] has two central characteristics. First, it is a physics in that it can be used to predict the qualitative behavior of devices. Envisioning is not concerned with *post-hoc* rationalization of observed behavior, but rather with constructing predictions that are consistent with observed device behavior. Second, it is a theory of causality in that it can be used to produce causal explanations acceptable to humans. 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 system by explaining how disturbances from equilibrium propagate. (Envisioning is often confused with qualitative simulation which it is only in its most degenerate form. In more complex cases it is primarily concerned with introducing and manipulating assumptions while maintaining a notion of causality.) A typical kind of physical mechanism we might envision is a pressure regulator (see Figure 1). A pressure regulator's 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): *"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 (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 remains approximately constant."*

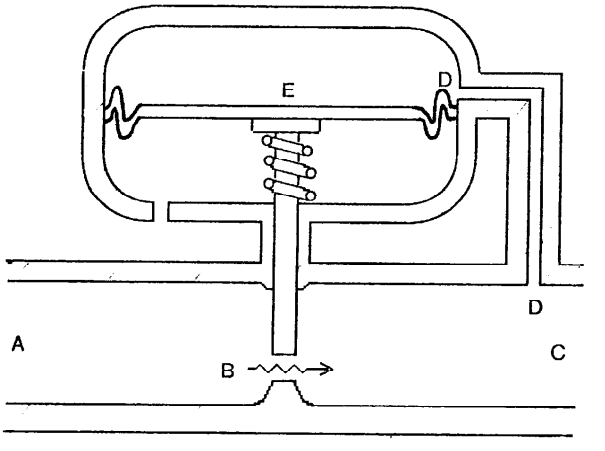

Figure 1 : Pressure Regulator

As this explanation for the pressure regulator illustrates, the task of envisioning is a difficult one. It must determine the causal inputs and outputs for each component, which can be a subtle task, especially when a component has more than two ports connecting it to other components of the device. It must detect and correctly determine the consequences of all the kinds 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]); rather, we examine the nature of the input evidence that the envisioning process operates on, the conclusions it produces and the relationship between these two.

STRUCTURE AND FUNCTION

An objective of the investigation is to explore a theory of causal reasoning that can, given a physical situation (in particular, a novel situation), correctly predict ensuing behavior in that situation. The situations 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 the situation is encoded? (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 interesting, or is most of the work it appears to be doing actually pre-encoded in the evidence provided to it?

One way to ensure that the "conclusions" have not been surreptitiously encoded into the evidence that the envisioner operates on is to make the evidence a well-defined and distinct ontological type, distinct from that of the conclusions. In particular, we require that the evidence be a description of the physical structure of the system, namely its constituent parts and how they are attached to each other. The conclusions describe the behavior, or functioning of the overall system. The task of causal reasoning is to deduce the functioning of the system 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 model library. Each type of component and connection has a specific model which describes its behavior in the abstract, independent of any 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 moving object are its position, velocity and acceleration. By modeling each component, the abstract qualitative behavior of the overall device is formalized as a set of qualitative equations. 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." The same library of models and the same causal reasoning process should successfully analyze a wide variety of physical devices, particularly devices that have not been analyzed before or devices under new operating conditions. We call this meta-theoretic constraint the *no-function-in-structure* principle. The class of devices constructable from any particular set of components is, in principle, infinite. We can only check our envisioning process on a finite subset of this infinite class and the no-function-in-structure principle provides some confidence that it will also succeed on the untested devices. As such the principle 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 switches for which current does not necessarily flow when they are closed (for example, two switches in series). Current flows in the switch only if it is closed and there is a potential for current flow.

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.

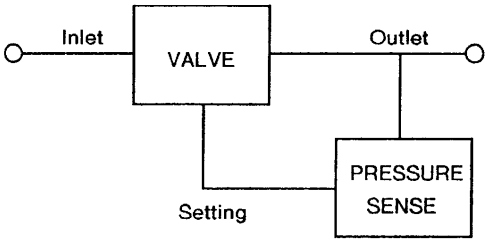

Figure 2 : Device Topology of the Pressure Regulator

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 system, the two attributes are volumetric-flow and pressure; for thermal systems, heat flow rate and temperature; for translational systems, force and velocity; for rotational systems, torque and angular velocity; for electrical systems, current and voltage.

COMPONENT MODELS

Envisioning is centrally concerned with qualitative incremental disturbances from equilibrium. This motivates the class of values (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 straight-forward, e.g., 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 qualitative equation (called a confluence): $AdP + PdA - dQ = 0$, where Q is the flow through the valve, P is the pressure across the valve and A is the area available for flow, and dQ, dA and dP represent disturbances from the equilibrium values of Q, A and P. In the situation where the pressure across the valve is positive (area is always positive) the expression simplifies to (using the qualitative calculus sketched out earlier): $dP + dA - dQ = 0$. 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 when it is completely closed.

The behavior of each state is provided by three types of rules. First, the model specifies the region of operation covered by the component state. For example, the closed state of valve is indicated by the condition $[A = 0]$, i.e., that there is no area available for flow. From these rules evisioning can determine what transitions between states are plausible. Second, the model provides confluences among a component's variables. These rules are used to determine what state a component might be in and to test whether a transition can occur. Finally, the model includes confluences among the changes in component variables. These confluences describe the incremental behavior of the component and are used in constructing causal explanations for device behavior.

The full model for the valve is:

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 straight-forward to identify the possible state transitions:

OPEN: $dA = -\Rightarrow$ WORKING

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

CLOSED: $dA = +\Rightarrow$ WORKING

THE ENVISIONING PROCESS

The envisioning process performs three kinds of analysis. It must determine (1) which state(s) the overall device is in, (2) the causal behavior of the device in each of those states, and (3) the possible transitions between the device states. Each type has a different form of explanation associated with it. However, in the remainder of this paper we will concern ourselves only with the causal behavior and its explanation (see [3] for more detailed discussions).

By modeling the behavior of each of the device's constituents, the potential behavior of the system is expressed as a set of confluences among changes in variables (e.g., dP, dQ). The type of the system (thermal, electrical, etc.) and the types of the variables (e.g., velocity, current, etc.) become irrelevant. Envisioning analyzes these confluences to construct a causal explanation for the behavior of the system.

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 predict correctly and satisfy no-function-in-structure, most are incapable of yielding any kind of reasonable explanation. For example, the best 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.

The explanations produced by envisioning are based on a very simplistic notion of causality which we call *naive mechanism*. A 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). That is, we do not seek a causal explanation of how it reached a given quiescent state or why it stays in that state but rather we seek a causal explanation of how the system responds to disturbances from a quiescent state. The difficult task for envisioning is constructing explanations within this limiting framework — constructing predictions alone is relatively easy.

ASSUMPTIONS AND PREDICTIVENESS

Because the information available to envisioning is qualitative, the actual behavior of the overall device may be underdetermined, i.e., more than one coherent behavior is 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 are subsequently reasoned upon. Thus, in underdetermined situations envisioning produces multiple interpretations, each with different assumptions and corresponding to a different overall behavior. At a minimum, for a prediction to be correct, one of the interpretations must correspond to the actual behavior of the real system. A stronger criterion follows from observing that a structural description characterizes a wide class of different devices. The prediction produced by an envisioning is correct if (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. Third, the notion of naive mechanism must be extended to include heuristic assumption steps in causal explanations.

CONCLUSION

Qualitative reasoning is a difficult task and ENVISION is a substantial program capable of producing interesting analyses for device behaviors that surprise even its implementors. In and of itself this last statement says nothing unless it also includes some assertions about the input evidence it operates on. A reasoning system should not be evaluated on the nature of its conclusions, but rather 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," Cognitive and Instructional Sciences, Xerox PARC, 1982.

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