SOME ISSUES ON MECHANISTIC MENTAL MODELS [1]

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

INTRODUCTION

Our long-range goal is to develop a model of how a person acquires an understanding of mechanistic devices such as physical machines, electronic and hydraulic devices, or reactors. We lay out a framework for investigating the structure of what we call *mechanistic mental models*: people's mental models of physical devices. Doing so involves developing a precise notion of a qualitative simulation. The concept of qualitative simulation derives from the common intuition of "picturing in one's mind's eye, how the machine operates."

Although one would intuitively expect qualitative simulations to be simpler than quantitative simulations of a given device, they turn out to be equally complex, but in a different way. These complexities arise, in part, from the fact that devices may appear nondeterministic and underconstrained when the quantities and forces involved in their makeup are viewed solely from a qualitative perspective. Therefore, if the qualitative simulation of the device is to behave deterministically, additional knowledge and reasoning must be used to disambiguate these "apparent" ambiguities.

It is surprisingly difficult to construct mental models of a device that are capable of predicting the consequences of events not considered during the creation of the model. Thus, the process for constructing a good mental model involves a different kind of problem-solving than the process for "running" the resultant mental model, a distinction that we find crucial for understanding how people use mental models. In fact, simply clarifying the differences between the work involved in constructing a qualitative simulation — a process we call *envisioning* — and the work involved in simulating the result of this construction — a process we call *running* — turn out to have both theoretical and practical ramifications.

QUALITATIVE SIMULATIONS

A Basis for Mechanistic Mental Models

Complex devices, such as machines, are built from combinations of simpler devices (components). Let us assume we know the behaviors of the components, as well as the way in which they are connected to form the composite device. The behaviors of the components are described qualitatively, such as "going up" or "going down," "high" or "low." The qualitative simulation always presents the events in the functioning of the machine in their causal order. Figure 1 illustrates a conventional door-buzzer (for the moment ignoring the button that activates the buzzer). The buzzer is a simple device, but complex enough to use for illustrating ideas of qualitative simulation.

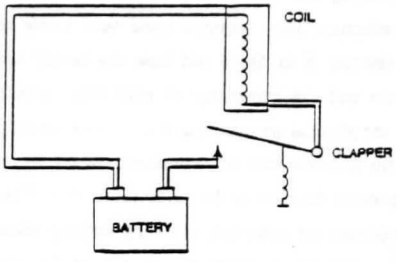

Figure 1 : Buzzer

The buzzer's qualitative simulation might be described as: *The clapper-switch of the buzzer closes, which causes the coil to conduct a current, thereby generating an electromagnetic field which in turn pulls the clapper arm away from the switch contact, opening the switch, shutting off the magnetic field, allowing the clapper arm to return to its closed position, and thereby start the whole process over again*[2]

The simplicity of the qualitative simulation as expressed in the preceding example is deceptive. Qualitative simulation encompasses a variety of ideas which need to be carefully differentiated. For example, we must distinguish simulation as a process from the results of that process. A simulation process operates on a representation describing the device, producing another representation that describes how the device functions. One source of confusion is that this latter representation can likewise be "interpreted" or simulated, but doing so will produce very little more than what is already explicitly represented in the functional representation produced by the first kind of simulation[3]

We need to distinguish four related notions which form the basic distinctions for a theory of qualitative reasoning. The most basic, *device topology*, is a representation of the structure of the device (i.e., of its physical organization). For example, the steam plant's structure consists of a steam generator, turbine, condenser, their connecting pipes, etc. The second, *envisioning*, is an inference process which, given the device's structure, determines its function. The third, *causal model*, describes the functioning of the device (i.e., a description of how the device's behavior results from its constituent components which is stated in terms of how the components causally interact). The last is the *running* of the causal model to produce a specific behavior for the device, by giving a chain of events each causally related to the previous one. Thus, both the structure and functioning of a device are represented by some knowledge-representation scheme

[1] This paper is an abridged and revised version of de Kleer & Brown [82].

[2] The repetitive opening and closing of the switch (i.e., its vibration) produces an audible sound.

[3] Note that this latter kind of simulation is just one of the kinds of inference mechanisms that can use or "interpret" the functional representation. Others can inspect it in order to answer such questions as "Could x cause y to happen?"

(device topology and causal model, respectively), with the former being the input to the envisioning process and the latter being its output; this output causal model is, in turn, then used in the running. The example of qualitative simulation presented earlier is ambiguous as to whether it refers to the envisioning, the causal model, or the running.

Envisioning, i.e., determining the functioning of a device solely from its structure often requires some very subtle reasoning. The task, in essence, is to figure out how the device works given only its structure and the knowledge of some basic principles. Structure describes the physical organization of the device, namely the constituent components and how they are connected, but it does not describe how the components function in the particular device. The "behaviors" of each component are described assuming nothing about the particular context in which the component is embedded (i.e., the description is context-free). These behaviors form a component model (or schema) which characterizes all the potential behaviors of the component; the envisioning process instantiates a specific behavior for each component from these models. These component models are the basic principles which the envisioning process draws upon to derive the functioning from the structure.

To determine the functioning of the overall device, each component's model must be examined and an individual, specific behavior instantiated for it. Thus, the functioning of the entire device is determined, in part, by "gluing together" the specific behaviors of all of its components. The problem for envisioning is determining for each component which behavior, given all the possible behaviors its model characterizes, is actually being manifested.

What makes the problem-solving effort involved in the structure-to-function inference process difficult is that the behavior of the overall device is constrained, not only by local interactions of its component behaviors, but also by global interactions. Therefore, in principle, the behavior models of the components which are specified qualitatively may not provide enough information to identify the correct functioning of the device. For example, if values are described qualitatively, often fine-grained distinctions cannot be made between them. Thus, in the case of the buzzer, the envisioning may not be able to determine which is greater, the force of the magnetic field or the restoring force of the spring. Knowing which is greater may, in fact, be crucial to deducing the correct functioning of the device.

In order to describe how the resultant behavior derives from the behaviors of the constituents, first, each important event in the overall behavior must be causally related to preceding events. Then, each causal relationship must be explained by some fragment of the component model of one of its components. The example describing Figure 1, is, at best, an abridged description of the buzzer's function. It causally relates each event to the preceding one, but fails to state any rationale for these causal connections. Because it is impossible to tell, a priori, whether the component models lead to unique behavior, the problem-solver must entertain the possibility that the structural evidence is underconstraining. Therefore the envisioning must take into account the possibility that one structure may have multiple possible functionings among which the envisioning cannot, in principle, distinguish.

"Running" the resulting causal model is closest to the original psychological intuition of "picturing, in one's mind's eye, how the machine operates." By running the model, one, in essence, does a straightforward simulation of the machine; the running itself does not have to determine or "prove" the causal or temporal ordering of events, as the envisioning process already has done so, and encoded the information in the causal model which serves as the input data for the running process.

The simplicity and elegance of the running process is the result of the complex problem-solving (i.e., envisioning) that constructed it. That our intuition that "picturing, in one's mind's eye, how the machine operates" is simple, is manifested by this running process. However, that sense of simplicity is deceptive, for the running is not possible without the more complex problem-solving which preceded it, removing all the ambiguities about how the machine *might* be functioning.

Understandably, the problems that arise in constructing causal models and the mechanisms that suffice in solving these problems are important for cognitive psychology and artificial intelligence. For psychology, they are important because they provide a framework for analyzing the "competency" involved in determining how a novel machine functions. Inasmuch as envisioning is restricted to being based solely on structural evidence, it becomes an interesting inference strategy in its own right for artificial intelligence applications, especially given the desire for artificial intelligence systems to be robust, and to be capable to deal with novel situations. The resulting models are more likely to be void of any implicit assumptions or built-in presuppositions based on how the device was intended to behave.

AMBIGUITIES AND ASSUMPTIONS

Origin of Ambiguities

In general, ambiguities originate from the fact that the information available to the qualitative analysis underdetermines or only partially characterizes the actual behavior of the overall device. There are three reasons for this underdetermination. The first and most obvious is that the quantities referenced by the component models are qualitative and thus fine-grained distinctions cannot be made between the attribute values or component states. Second, because the implicit time progression in the simulation is qualitative, it is not always possible to determine the actual ordering of events. And the third reason, not directly related to the qualitative nature of the models, comes from the limitations on the kinds of information captured by the models. Because envisioning tries to identify a global flow of action by piecing together local cause-effect rules of the component models, a component model encodes only those aspects of the component's behavior that can be used in such a fashion. However, our understanding of a given component often involves more knowledge than is (or, perhaps, could be) encoded in such mechanistic rules. For example, in modeling the internal operation of a pump we know from the laws of physics that fluid is conserved in passing through the pump. But, because this piece of knowledge is a *constraint*, it cannot be represented by

any cause-effect rule: the inability to encode it can lead to a given component model being underdetermined.

Origin of Assumptions

In the buzzer example, because of the qualitative nature of the attribute values, the envisioning process cannot determine whether the spring is stronger than the magnetic field. In this "impasse," it is forced to consider two hypothetical situations: one in which it *assumes* the spring is stronger than the magnetic field and one in which it *assumes* the spring is weaker than the field.

Impasses occur when envisioning cannot evaluate a transition condition (e.g., the condition of the switch being open) or invoke an attribute equation (e.g., that of field strength being proportional to coil current) to determine the value of an unknown attribute. In order to proceed around impasses, the envisioning must introduce assumptions about the truth or falsity of conditions or about the values of unknown attributes.

The buzzer example can be used to illustrate an impasse which arises from the envisioning being unable to determine whether a transition condition holds. In this impasse, the envisioner introduces an assumption that the condition "force from the coil > restoring force of the spring" is true, and then proceeds to analyze the new resulting state. Of course, the resulting causal model will then contain two accounts of the device's functioning: one in which the clapper rises and one in which it does not. Additional knowledge and reasoning strategies must then be used to verify or reject the various assumptions that were created to enable the envisioner to proceed around such impasses. These strategies combined with a much more extensive analysis of the kinds of assumptions needed in order to construct a causal model have been detailed in the expanded version of this paper.

REFERENCES

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.