9 Mental Models of Physical Mechanisms and Their Acquisition

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

INTRODUCTION

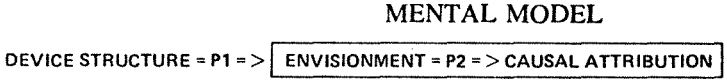
In the past, many research projects have aimed at a better understanding of how one learns a language, a set of procedures, a large corpus of facts, nonsense syllables, etc. In this chapter, we focus on a new domain—one that shares very few properties with any of the previous. The domain is that of mechanistic devices, including physical machines, electronic and hydraulic systems, and even hybrids such as electro-mechanical systems. Our top-level goals are: (1) to investigate what it means for a person to understand a complex system, in particular, the mental models that experts form of how a system functions given the system's constituents and their interconnections; and (2) to discover principles that orient the acquisition of the capability to construct these models.

It is no surprise that these two goals are interrelated. In fact, it is nearly impossible to investigate the properties of the acquisition process without first having a clear notion of exactly what must be learned. Thus, we see that the first goal is a prerequisite for the second. What is surprising is that the second goal is intimately tied to the first, in that the properties that make these mental models learnable or that facilitate their development turn out to be essential characteristics of a highly robust mental model in the first place. In fact, each of the principles we discuss for guiding the learning process are equally well motivated as principles for guaranteeing the robustness of the mental model. As we shall see, the duality of these principles is not accidental. This means that our chapter can be read from two quite different perspectives. The first concerns what constitutes mental models of mechanistic systems and what characteristics enhance their utility for problem solving by an expert. The second concerns how these characteristics impact the acquisition of skills for constructing mental models.

Before we begin a technical discussion of these issues, we need to provide an overview of the kind of mental models we are discussing here, the processes underlying their construction, and the kind of learning principles that we are after.

The kind of mental models of a mechanistic system that we are interested in are generated, metaphorically speaking, by running a qualitative simulation in the mind's eye. We call the result of such simulation "envisionment." One of the most important properties of envisionment is its ability to manifest a system's causality, which not only makes it extremely useful for constructing causal models of how and why the system functions, but also makes the envisionment sufficiently self-evident that it, also, can be "run" efficiently in the mind's eye; that is, envisionments have the property that each new state change is *directly* caused by a prior event. Hence determining the next state involves very simple reasoning (de Kleer, 1979).

We can portray the basic entities involved in constructing mental models in a diagram, such as the one below:



P1 denotes the process of examining the device's structure and from that constructing an envisionment of its behavior. P2 denotes the process of examining the resulting envisionment—both its form and its execution—and from that deriving a causal attributive description. The mental model is a combination of an envisionment and its causal attribution. But it is not a fixed entity because the attribution process (P2) can either be run to completion producing a complete description or can be allocated limited resources producing a partial description. For the purposes of this chapter, the details of the two processes are of secondary interest. Instead, our attention is focused primarily on discovering a set of principles of critiquing candidate mental models.

Briefly, it is the task of the learning system to provide an adequate process for generating an envisionment from a structural description of a device knowledge of what the device is supposed to do. Learning how to generate adequate envisionments is exceedingly complex, and we do not have a complete description of the learning process at this point in our research. However, the learning task falls naturally into two phases. It is the responsibility of the first to critique the given envisionment and isolate any parts not conforming to the behavior of the physical device or not meeting certain principles. By striving to meet these principles, an *esthetic* for the learning process is provided, thus setting the stage for the learning system to improve its current envisionment, even if it is already accurate and consistent. It is the responsibility of the second phase to critique the model construction process to isolate responsibility for the critiqued envisionment.

THE MOTIVATION FOR DEFINING SOME ESTHETIC PRINCIPLES

The purpose of defining a set of esthetic principles for critiquing simulation models—even those that are already consistent—is to provide a gradient or direction of progress for the learning process to improve its current model of the system. But toward what end?; that is, what are the properties of an ideal envisionment, and why are these properties important? The answer lies in the essential properties of an expert's models, those properties that maximize his ability to use the model to answer unanticipated questions or predict the consequences of novel situations.

Our approach to discovering these properties has not been to probe the models of an expert directly but rather to hypothesize what such a model might be, criticize it according to principles abstracted from our experience with its implementation, and then characterize the essential properties of the model that maximize its robustness (Brown, Collins, & Harris, 1977).

THE NO-FUNCTION-IN-STRUCTURE PRINCIPLE

We assume that a given system consists of a set of components, each of whose behaviors are satisfactorily modeled by a set of rules, and expect our simulation or envisionment to explain how the system's overall behavior is caused by the behavior of its constituents and their interactions. The preceding process is potentially recursive in that models can be embedded in models. Said differently, the behavior of a system's component can be characterized by a set of rules at one level of detail, and at another level be viewed as a system whose behavior is, itself, derivable from still lower-level components. Our concern is not the choice of appropriate level of modeling but rather, at a given level of detail, to examine the problem of constructing a qualitative simulation that is maximally robust. This robustness is achieved by ensuring that the system's simulation model does not presuppose the very mechanism it is trying to describe. We do this by having it satisfy a constraint called the "no-function-in-structure" principle. This principle states that the rules for specifying the behavior of any constituent part of the overall system can in no way refer, even implicitly, to how the overall system functions.

Violating the no-function-in-structure principle also limits one's ability to use the simulation in predicting how the system might function, given a modified or faulted component. Because a slightly modified component can sometimes radically alter the underlying mechanism of the overall system, those rules that specified a component's behaviors by implicitly presupposing this mechanism may no longer be valid. The predictive aspects of the simulation in such a case are highly questionable (Brown and Burton, 1975).

Fully satisfying this principle ensures that the behaviors of each of the system's parts can be represented and understood in a context-free manner, independent of the overall system. Failing to adhere to this principle leads to an explanation of how the system functions predicated on some aspect of how it functions. This recursion is not always bad; often, on first explaining a system, it is best to violate this principle, providing an explanation that establishes a framework on which to hang more satisfactory mental models.

THE WEAK CAUSALITY PRINCIPLE

Whereas the main purpose of the no-function-in-structure principle is to establish a criterion by which to evaluate hypothetical models, there are other principles that facilitate the learning process. The weak causality principle is the most important of this class. It states that every event must have a direct cause, meaning that the reasoning process involved in determining the next state change is local and does not use any indirect arguments in determining what the next state must be. An envisionment that satisfies this principle has the useful property that one can precisely identify which aspect of the envisionment is responsible for any piece of behavior that is inconsistent with the observed device. In other words, a learning system following this principle can discover where to assign credit or blame in the current model.

Duality of Purpose

Seen from the perspective of learning, the weak causality principle facilitates the assignment of credit. It is interesting to note that this same principle has quite a different role when seen from the perspective of performance—that is, performing the problem solving needed to construct the envisionment. From this perspective, a model that satisfies this principle enables the set of next possible states to be determined without requiring elaborate reasoning. Constraining the weak causality principle to require that the next state is also uniquely determined has a major impact on the efficiency of constructing the envisionment as well as facilitating the ease of using it. Later in the chapter we expand this strong causality principle.

The duality also holds for the no-function-in-structure principle. The latter provides a direction for improvement when used in a learning context, and it provides a guarantee of robustness when used in a problem-solving context.

METHODOLOGY

Our approach has been to study a class of devices that on the one hand are intuitively understood by nearly everyone but on the other hand are sufficiently

complex to raise many problematic issues—some of which were unanticipated by us. The device we have chosen is a simple doorbell or electro-mechanical buzzer. It is sufficiently complex to stress any current theory of causal mechanisms including our own. Finding an internally consistent qualitative simulation of the buzzer that also satisfies both the causality principle and the no-function-in-structure principle turns out to be surprisingly difficult. Similarly, because the buzzer is a common household device, the reader can judge the subtleties involved in finding a principled simulation and appreciate how that simulation can be used to answer a wide variety of questions. Granted that the need for defining a set of principles constraining the representation of a system might have been more convincing had we chosen a more complex paradigmatic example, the chance of readers following and even anticipating the various pitfalls that unfold as we explore a sequence of potential models would have been radically reduced. Again, let us stress that, like any single device however complex, the buzzer itself is of no fundamental importance.

TECHNICAL ISSUES

What form should such a representation take, and what should we expect to be able to do with it? As with much of this chapter, let us proceed by example. Pictured here is a diagram of the buzzer. Its functioning seems to be easily

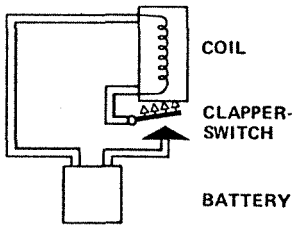


FIG. 9.1. Buzzer.

described by a small set of rules basically asserting: *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, thereby opening the switch, which shuts off the magnetic field allowing the clapper arm to return to its closed position, which then starts the whole process over again.* Indeed, we could easily create a set of formal rules that would produce this ad hoc description, but does that description produce a useful understanding?

This question can be answered using the following definition of mental model robustness: A model is robust with respect to its device structure, if the questions that can be asked about the device structure can be answered correctly. The

device structure implicitly defines the terms of a descriptive language, and questions using those terms concern the ramifications of changing the device structure or its component attributes. Our ad hoc description of how the buzzer is supposed to buzz does not provide an understanding that can address such possible questions.

For example, listed here are some typical questions that one might expect a person who has a thorough, qualitative understanding of the buzzer's mechanism to be able to answer without recourse to analytic models of its components, nor to differential equations, etc:

1. What happens if we reverse the leads of a battery (i.e., change the polarity of the battery in the circuit)?
2. What happens if we switch the leads of the coil?
3. What happens if we remove the battery from the circuit?
4. What happens if we short the switch contact?
5. What happens if we make the switch arm lighter (or even weightless)? For example, what happens qualitatively to the frequency of the vibrator or the size of the gap when such a change is made?
6. What happens if we place a light-weight permanent magnet on the clapper arm? Does the frequency of the buzzer change? Does it matter which way the battery is hooked up?
7. What happens if we put two clapper switches in series (or parallel)? If one buzzes with a lower frequency than the other when subjected to the same magnetic field, then what happens when they are placed in series?

Attempting to answer questions like the preceding—some of which are admittedly quite hard—demonstrates the surprising richness of inferences that can follow from a totally qualitative understanding of the underlying mechanisms of the buzzer. Similarly, the inadequacies of the previous description of how the buzzer functions become quite obvious. To answer these and other unanticipated questions just from a representation of how the buzzer works places quite a burden on that representation and the processes that interpret it. But apart from the failure of that description to answer such questions, there is a principled objection: A great deal of that description already presupposes how the buzzer works.

For example, the statement, "the clapper switch closes causing the coil to conduct a current" presupposes a source of current or battery in the circuit. Furthermore, it presupposes that the switch, coil, and battery are all arranged in a very particular way, and that when the switch closes it necessarily passes current. Switches in general do not pass current; they only pass current if they are connected to batteries and wires in particular configurations. Although it may be true for this buzzer that the switch passes current, we can only say so because we already understand how it works.

The desired level of understanding is this: Given a description of how the switch operates in all possible circuits, use that along with similar descriptions of the other parts of the buzzer to explain how and why the switch passes current as it closes. In general, the problem is how to connect the functioning of the device (i.e., its buzzing) with the structure of the device (i.e., the buzzer) without presupposing how the device functions.

Our distinction between structure and function is crucial, even in any formal analysis of the buzzer. For example, the coil could be modeled by $V = L(di/dt)$—a model valid for any electromagnet—and such models could be combined into differential equations and solved in standard ways. But for the engineer to assert a priori that the voltage V across the coil necessarily obeys some particular function, f(t), would render his analysis suspect, because he would be using the solution in order to find it. Without an appropriate distinction between structure and function, an analysis can be uninformative. We assert that a meaningful analysis based on either quantitative or qualitative component models must not allow any of the solution (i.e., function) into the description (i.e., structure) of its pieces.

In the following sections, we present a series of successively more adequate simulation models. Each of these models is of interest in its own right in that each illustrates a problematic issue in constructing or critiquing possible envisionments.

A SIMPLE QUALITATIVE MODEL FOR THE BUZZER

The fundamental question we want to address is, given that the behaviors of components (magnets, batteries, coils, etc.) are understood generally and sufficiently, how do the understandings of the behaviors of the individual components combine to explain the behavior (buzzing) of the composite device (buzzer) constructed from these components? The following is an extremely simple model for describing how the buzzer functions. The purpose in presenting this model first is to introduce our techniques for formalizing qualitative models, illustrate the difficulty modeling even a simple device, and demonstrate that a formal qualitative model that does not obey certain principles is not necessarily very useful.

The two critical components of the buzzer are the coil, which produces a magnetic field, and the clapper switch, which moves in this field. The clapper switch is either closed, connecting the battery and the coil, or open, disconnecting them. The coil is either on, causing a magnetic field that attracts the clapper switch, or off, causing no field. The behaviors of the two components can be formalized as:

SWITCH : OPEN: battery and coil are disconnected.
CLOSED: battery and coil are connected.

COIL : ON: magnetic field is pulling at switch.
OFF: no magnetic field is pulling at switch.

Each component of the buzzer is indicated and its possible states are distinguished. For each such state, a description of its behavior in that state is included.

This model is incomplete because it does not indicate how the magnetic field of the coil can change the state of the switch, nor how the battery can change the state of the coil. We need to somehow associate pulling at the switch with its being open, and the battery-to-coil connection with the coil's being on. One possibility is to extend the description of the behaviors:

SWITCH : OPEN: battery and coil are disconnected,
magnetic field is pulling at switch.
CLOSED: battery and coil are connected,
no magnetic field is pulling at switch.
COIL : ON: magnetic field is pulling at switch,
battery and coil are connected.
OFF: no magnetic field is pulling at switch,
battery and coil are disconnected.

The rationale behind this model is that if the switch is open, the magnetic field must be pulling at it, because the field is what opens it. Similarly, if the coil is on, the battery must be connected to it, because otherwise the coil couldn't produce a magnetic field.

In order to understand this model's implications, consider a particular situation in which the switch is closed:

SWITCH : CLOSED: battery and coil are connected,
no magnetic field is pulling at switch.

According to the preceding model, the coil can now be neither on nor off. If the coil is off:

COIL : OFF: no magnetic field is pulling at switch,
battery and coil are disconnected.

The coil rule "battery and coil are disconnected" contradicts switch rule "battery and coil are connected." Further examination shows that every other possible component state is similarly contradictory. The model is completely inconsistent and thus of little use for understanding the operation of the buzzer.

A TIME MODEL OF THE BUZZER

The difficulties of the previous models cannot be removed by simply changing the descriptions; any model for the buzzer of the previous form will be inadequate, because the form has failed to consider time and causality. For instance

"battery and coil are disconnected" does not necessarily contradict "battery and coil are connected," if these two hold at different times. By modifying the form to distinguish definitional from consequential behaviors, a qualitative notion of time can be introduced into the buzzer model. The symbol "−>" is used to indicate consequential behavior.

SWITCH : OPEN: battery and coil are disconnected,
−> no magnetic field is pulling at switch.
CLOSED: battery and coil are connected,
−> magnetic field is pulling at switch.
COIL : ON: magnetic field is pulling at switch,
−> battery and coil are disconnected.
OFF: no magnetic field is pulling at switch,
-> battery and coil are connected.

The rationale for this switch model is that if the switch is open, the magnetic field will soon be cut off, because the battery is disconnected. Similarly, if the coil is on, the battery will soon be disconnected, because the magnetic field opens the switch. Because this model has an explicit notion of time, it is possible to simulate the behavior of the buzzer model over time. One way of describing the result is to arrange in tabular format the states of the components at successive time increments (we discuss this process in more detail in the next section):

t	SWITCH	COIL	BATTERY	->BATTERY	COIL	->COIL
0	CLOSED	ON	connected	disconnected	pulling	pulling
1	[OPEN]	ON	disconnected	disconnected	pulling	not pulling
2	OPEN	[OFF]	disconnected	connected	not pulling	not pulling
3	[CLOSED]	OFF	connected	connected	not pulling	pulling
4	CLOSED	[ON]	connected	disconnected	pulling	pulling
5						

The boxes indicate which states changed from the previous time increment. In the initial state (t = 0), the coil being on results in the battery being disconnected; therefore, the switch must change state from closed to open. In the next state (t = 1), the fact that the switch is open results in the coil not pulling; therefore, the coil must change state from on to off.

Although this model is internally consistent and capable of explaining the phenomenon (the vibration), it does not represent a very robust understanding of the buzzer. First, the model for the coil inherently assumes its connection to a battery and a switch, and, furthermore, it assumes that the battery-switch connection allows the battery to be disconnected from the coil whenever the magnetic field is pulling at the switch. This particular model for the coil has the entire functioning of the buzzer embedded within it and thus is inadequate for any device except this particular buzzer. It violates the no-function-in-structure principle. Second, the successor state is determined indirectly. For example, the consequence of the coil's being on is the battery's disconnection, and the only

consistent state for the switch, in this case, is open. This indirect determination of state makes it impossible to determine credit or blame. The weak causality principle has been violated.

A CAUSAL MODEL FOR THE BUZZER

The difficulties of the previous buzzer models force us to expand the syntax to conditionalize the consequences and to make a component's consequences affect only the component itself. This change addresses both the no-function-in-structure principle and the weak causality principle. Because the model for a component is now less dependent on its surrounding context, more robust models for the components can be constructed. The conditional makes it possible to identify precisely why a component changes state, making it is easy to assign credit or blame.

The rule for a state consists of a definitional part (e.g., the battery is disconnected), and a conditional that makes a test (e.g., if the coil is not pulling) to determine whether some consequence applies (e.g., the switch will become closed). The general form for a component model is:

<component> : <state1>: (<definitional-part>)*, (if <test>,
<consequence>)*
<state2>: (<definitional-part>)*, (if <test>,
<consequence>)*

"(...)*" indicates that "..." may occur an arbitrary number times. The <definitional-part> can be used in two ways. The first concerns its use as a criterion for determining whether the component is in a given state, and the second concerns its use as an imperative. Given that the component is declared to be in a particular state (criterion), then statements made in the <definitional-part> are asserted to be true (imperative) about the component's behavior. These assertions can then be examined by an inferential process to determine their ramifications. In simple cases, these ramifications can be determined by examining the tests of the current state of every component model.

The models for the buzzer parts are:

SWITCH : OPEN:
battery is disconnected,
if coil is not pulling, switch will become CLOSED.
CLOSED:
battery is connected,
if coil is pulling, switch will become OPEN.
COIL : ON:
coil is pulling,
if battery is disconnected, coil will become OFF.

OFF:
coil is not pulling,
if battery is connected, coil will become ON.

In order to combine these descriptions of the behaviors of the individual parts to determine the behavior of the overall buzzer, we set out to construct a simulation. Starting with some arbitrary[1] initial state for each of the components (i.e., a cross product or composite state), their definitional parts are asserted, and then each test is examined to determine the consequences of those assertions being true. From these consequences, the next state is determined. The discovery of each next state can be considered as a time increment. The time elapsed between increments is variable, because increments refer to interesting events rather than the passage of some fixed unit of time. The determination of the next composite state can be very complex. The consequences of the prior composite state may indicate that multiple components will change state in the next time interval. In the actual physical system, it may be critical that one component change state before another, but in a qualitative simulation, the time scales for each of the models are not inherently comparable, because they don't utilize a uniform time metric. The exact ordering of the transitions cannot be directly determined without considering the overall purpose of the device and other nonlocal properties that require complex inferencing schemes. A second complication stems from the qualitative nature of the inputs and other internal parameters. In this case, it might be ambiguous whether some threshold in the test was passed, resulting in various components optionally changing state. Such a situation can force the consideration of parallel envisionments.[2]

The transition table is:

t	SWITCH	COIL	BATTERY	COIL
0	CLOSED	ON	connected	pulling
1	[OPEN]	ON	[disconnected]	pulling
2	OPEN	[OFF]	disconnected	[not pulling]
3	[CLOSED]	OFF	[connected]	not pulling
4	CLOSED	[ON]	connected	[pulling]
5				

In the initial state (t = 0), the coil's being on causes the switch to change state from closed to open. If the switch is open, the battery is disconnected by defini-

[1]For more complex devices, this choice cannot be arbitrary, because some of the composite device states may be either contradictory or inaccessible—requiring several initial states. A composite state is contradictory if the definitional parts of two of its component states make contradictory assertions.

[2]However, many of these difficulties can be avoided by detecting and eliminating self-contradictory device states and invoking other deductive machinery. It is the responsibility of the simulation process P1 to identify the ambiguities, but it is also the responsibility of P1 to prune as many of the resulting envisionments as possible based on local considerations.

tion. In the next state (t = 1), the battery's disconnection causes the coil to change state from on to off. If the coil is off, it is not pulling by definition. The construction continues. The simulation can also be represented as a state transition diagram.

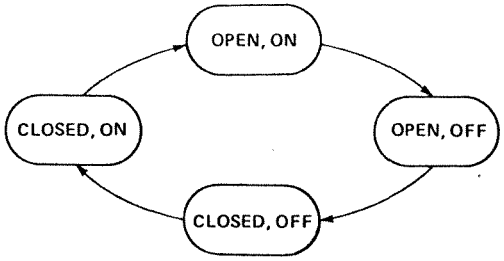


FIG. 9.2. State-transition diagram for the buzzer.

The preceding table can also be used to construct a series of snapshots of the buzzer functioning over time (see Fig. 9.3).

NO-FUNCTION-IN-STRUCTURE AND THE LOCALITY PRINCIPLES

The buzzer model, although it explains the vibration, has so much of the knowledge of how the buzzer works embedded in the models of the individual components that it provides little insight into how the buzzer, as an interacting collection of components, functions. Implicit within each of the models is that it is connected to other constituent components comprising a buzzer. The models name other components and their internal states directly, thereby presuming that they are part of a buzzer. That the components are physically connected by wires and magnetic fields is often ignored. For example, the model for the switch assumes its connection to a battery and that when it opens it will prevent current from flowing and disconnect the battery from the coil. The model for the coil assumes that it is not the current flowing through it but the fact that some battery is connected to something, which enables the magnetic field to exist. Thus, much of the essential functionality of the overall buzzer has been embedded within its constituent parts. There is little chance that these same models would be useful for understanding a buzzer that was constructed even slightly differently. For example, a buzzer having two clappers hooked in series or parallel is sufficiently different for these models to be of little use.[3]

[3]Notice that in this case, contradictory states exist: If one clapper is open and the other is closed, the rules make contradictory assertions about whether the battery is connected. Or if that were resolved by having better switch models, the next state of the coil is ambiguous.

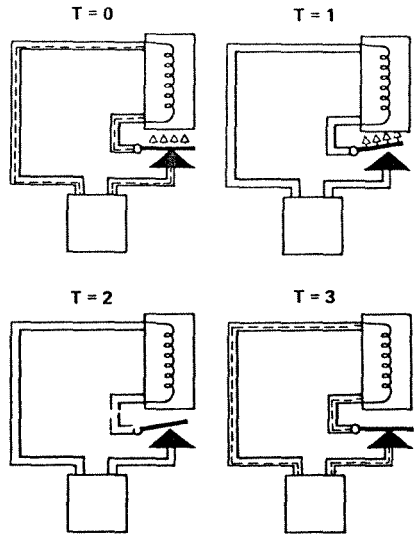


FIG. 9.3. The functioning of the buzzer.

Syntactic restrictions can be placed on the models of the individual components to ensure that they reference *local* quantities. Such locality restrictions help to avoid gross violations of the no-function-in-structure principle. The first locality principle demands that there is no reference to other components in rule consequences. The previous models' rules met that principle. A similar locality principle can be used to restrict the definitional aspect of component rules. But if a locality principle were also enforced on the tests of component rules, there would be no way for models of different components to ever interact; thus, there is no way to avoid some nonlocality. To extend the locality principle to all three parts of a rule forces the introduction of connections as entities independent of the component models.

CONNECTIONS AND DEVICE TOPOLOGY

In order to avoid nonlocality in the component models, we need to draw a distinction between the models for components and the method by which these models communicate. We introduce *connection* as a simple stateless model that is primarily a placeholder for information. Although we make the internal state of a component inaccessible to other components and connections, models for components and connections communicate by sharing information. For example, both the model for a specific wire (e.g., wire 2) and its adjacent part (e.g., coil) will know the current flowing from the wire into the coil. The simplest model for a wire is one that consists of the knowledge of the current through it, and it shares this information with the components on either end of the wire. The only infor-

mation that is shared by connections are *attributes* that are related to the actual physics by which components interact (e.g., voltage, current, pressure, force, flow, etc.) The no-function-in-structure principle also applies to connections. For example, it requires every wire in the buzzer to be modeled in the same way, and also every attribute of the same type (e.g., force) to be treated in the same way (e.g., obey Newton's laws). The component–connection distinction allows us to model the effect of the coil on the clapper switch: The coil being on (state of a component) causes a strong magnetic flux (attribute of a connection) that causes the clapper switch to open (state of a component).

Formally, it can often be arbitrary as to which parts of the buzzer are components and which are connections. All the buzzer models presented so far implicitly assume wires and magnetic fields to be connections. This is not necessary; we could have modeled a wire as a component having certain properties, but then we would have had to introduce an end-of-wire connection to attach the wire component to the switch component. The determination of which parts should be modeled as connections and which as components can be quite subtle, because a connection assumes that the model describing its behavior is extremely simple. For example, it would have been, in principle, impossible to model the switch as a connection because a switch has internal state. If we had decided the switch was a connection, we would never be able to construct an adequate model for the buzzer.

The device topology of a physical device consists of a description of the components of the device and their connections.

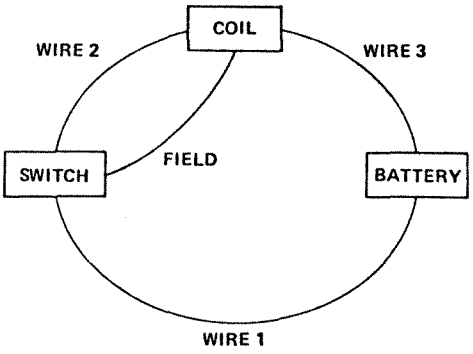


FIG. 9.4. Device topology of the buzzer.

Although a specific model for a specific component is permitted to know about the current in a specific wire, the class wide or prototype is not. Therefore, the prototype model can only express its potential connection to some wire. It refers to information that it might possibly share with connections as variables. When the prototype model is used in a specific device, these variables need to be replaced with device-specific quantities, because that is the only way two component models can communicate.

A prototype model for the switch is:

Information terminals: i1, i2, f1
SWITCH : OPEN:
i1 <− 0, i2 <− 0
if f1 = 0, SWITCH will become CLOSED.
CLOSED:
i1 <− 1, i2 <− 1
if f1 = 1, SWITCH will become OPEN.

A variable such as "i1" is intended to indicate the value of some attribute (e.g., a current of one ampere) that can then be communicated among models. Unlike component states that do not change until acted upon by some other model, variable values are direct results of components being in particular states and thus are only considered valid as long as the components that caused them do not change state. If a variable value is changed by some model, it may not be changed again until the component that originally caused the variable value changes state.[4] The semantics of "a <− b" is a[t+1] = b[t] for each time t in which the rule is valid. Or informally, a's value gets b's value. In the previous model, b is a constant, and the effect of "a <− b" is to set a's value to that constant. For example, as long as the switch is open, the rule "i1 <− 0" applies, thus i1 is set to zero immediately after the switch becomes open and cannot change until after the switch ceases to be open. The actual amount of time elapsed moving from t to t+1 is arbitrary—it can be infinitesimally small or extremely large. Our convention is that time t+1 refers to the next interesting event after time t, thus the time elapsed moving from t to t+1 has no relation to the time elapsed moving from t+1 to t+2.

A prototype model for the coil is:

Information terminals: i1, i2, f1
COIL : ON:
f1 <− 1
if i1 = 0, coil will become OFF
if i2 = 0, coil will become OFF.
OFF:
f1 <− 0
if i1 = 1, coil will become ON
if i2 = 1, coil will become ON.

An overly simplistic prototype model for the battery is:

Information terminals: i1, i2
BATTERY : i1 <−> i2.

[4]If this is violated, for example, if one component changes the current to one ampere and another to two, the model for the composite device is inconsistent.

"a <-> b" is a short-hand for "a -> b and a <- b." In other words, propagate changes in whatever direction the behavior of the device dictates.

We now have a well-defined, precise way of moving from a device topology to a set of device-specific models that does not violate the no-function-in-structure principle. For each node in the device topology where a connection attaches to a component, a new unique quantity must be invented; then for each component and connection, a copy of its prototype is made with the information terminals replaced with the appropriate circuit-specific quantities. This process ensures that the component models are local, because the only quantities a model can reference are those that are associated with the connections that are adjacent to it in the device topology. Thus, many of the violations of the no-function-in-structure principle are avoided.

One possible set of specific models for the buzzer are:

```
SWITCH : OPEN:
            I1 <- 0, I2 <- 0
            if F1 = 0, switch will become CLOSED.
         CLOSED:
            I1 <- 1, I2 <- 1
            if F1 = 1, switch will become OPEN.
COIL : ON:
         F1 <- 1
         if I2 = 0, coil will become OFF
         if I3 = 0, coil will become OFF.
       OFF:
         F1 <- 0
         if I2 = 1, coil will become ON
         if I3 = 1, coil will become ON.
BATTERY: I1 <-> I3.
```

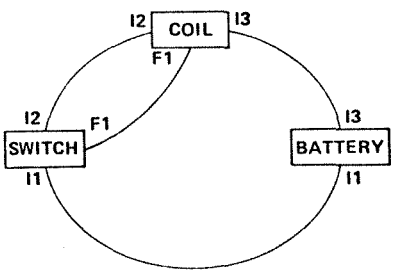


FIG. 9.5. Model for buzzer.

We arbitrarily chose names for the circuit quantities I1, I2 over QUANTITY1, QUANTITY2 purely for expository reasons. The models for all three parts are completely symmetric, and it does not matter which way these components are placed in the buzzer. The simulation constructed using the previous models is:

t	SWITCH	I2	COIL	F1
0	CLOSED	1	ON	1
1	[OPEN]	1	ON	1
2	OPEN	[0]	ON	1
3	OPEN	0	[OFF]	1
4	OPEN	0	OFF	[0]
5	[CLOSED]	0	OFF	0
6	CLOSED	[1]	OFF	0
7	CLOSED	1	[ON]	0
8	CLOSED	1	ON	[1]
9				

The models used for the wires are not really adequate for more complex circuits, and a better model would have been one in which the wire was modeled as having two ends where the currents in the two ends were related by i1 <−> i2. In such cases, the connections would be viewed more as conduits that have two ends with laws determining the transmission of information through the conduits. For the buzzer, this would only have obscured the point we were trying to illustrate.

A MYTHICAL TIME MODEL

The previous model has a serious problem with time. Doing the analysis to construct the entire table we see the problem:

t	SWITCH	I2	COIL	F1	I1	I3
0	CLOSED	1	ON	1	1	1
1	[OPEN]	1	ON	1	1	1
2	OPEN	[0]	ON	1	[0]	1

There are potential difficulties in the last two rows of the table. The current on one side of the coil I2 is 0, whereas the current on the other side of the coil I3 is still 1, because it takes one more time increment until I3 becomes 0. First, this raises the question whether the coil should really change state—although I2 = 0 is evidence for changing the coil's states from on to off, I3 = 1 is evidence for immediately changing the coil's back to on. Second, it violates Kirchoff's Current Law that requires the current flowing into a component to be equal to the current flowing out of it. We need to distinguish between the time taken for a component to change state and the instantaneous propagation of information in the connections or components that do not change state. For example, it is a myth that the flow of current in the wires happens instantaneously. We call such instantaneousness "mythical time," because no time elapses from the point of view of the state transitions, and the only reason it is important to retain some notion of time, albeit mythical, in the connections is to establish a causal rela-

tionship among the changes in attribute values (in this case current).[5] If all possible mythical time increments are considered before considering the next real time increment, the difficulties are resolved. Mythical time is indicated by hyphens in the transition table.

t	SWITCH	I2	COIL	F1	I1	I3
0	CLOSED	1	ON	1	1	1
1	[OPEN]	1	ON	1	1	1
1-1	OPEN	[0]	ON	1	[0]	1
1-2	OPEN	0	ON	1	0	[0]
2	OPEN	0	[OFF]	1	0	0
2-1	OPEN	0	OFF	[0]	0	0
3	[CLOSED]	0	OFF	0	0	0
3-1	CLOSED	[1]	OFF	0	[1]	0
3-2	CLOSED	1	OFF	0	1	[1]
4	CLOSED	1	[ON]	0	1	1
4-1	CLOSED	1	ON	[1]	1	1
5	...					

The information in the preceding table can be represented in an expanded state transition diagram. First, every change in a value or state is represented by a node (indicated by a boxed entry in the table). Second, edges are drawn from each state or value back to the states and values that caused them (these can be determined by examining the component rule that computed the new value or state). Arrowheads are placed pointing toward the new value or state. Third, the nodes are grouped together in larger nodes according to elapsed time increments. The resultant diagram is Fig. 9.6.

The behavior of the buzzer can now be regenerated by just tracing along the edges in the direction of the arrowheads.

A REPRESENTATION FOR CAUSAL ATTRIBUTION

Rieger and Grinberg (1977) have developed a system for representing cause-effect knowledge about physical mechanisms. A significant contribution of their work is an epistemology for the functionality of a physical mechanism. For example, their representation distinguishes between a state change and a tendency to produce a state change and allows a tendency to cause a state-change but not vice versa. However, the state of an object may enable a tendency to exist that causes a state change. This same effect is manifested in our models: The only way one part can affect another is through connections.

[5]The notion of attributes will become more important when the stuff flowing in the conduits is modeled in more detail, in that the stuff can have multiple attributes (e.g., voltage and current).

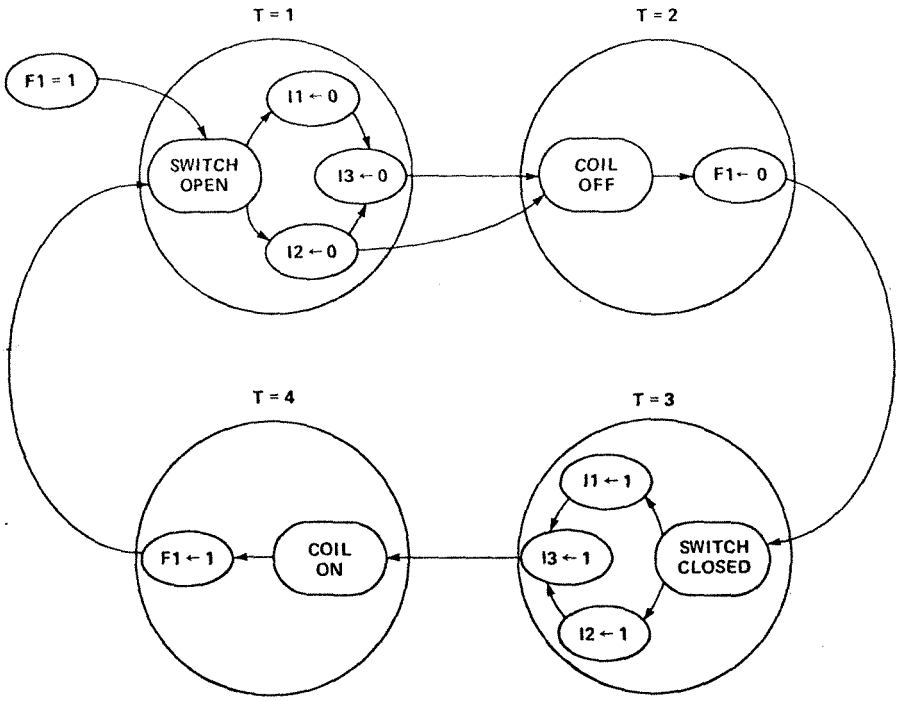


FIG. 9.6. Causation in the buzzer.

The prior analysis of the buzzer can be used to construct a crude representation similar to Rieger's. Events are represented by nodes of which there are only two types: (1) state changes, which represent a change in a component state; and (2) tendencies, which represent attributes of the connections being forced to some value. There are four types of links that represent the causal relationships between these two types of events.

The first link is **enablement:**

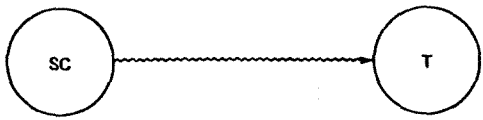


FIG. 9.7. Enablement.

The tendency T is a direct result of some component changing to a particular state SC, because the component being in that state SC is what enables tendency T to exist. For example, if the switch changes to the open state, the tendency of the current is toward zero.

The second link is **cause:**

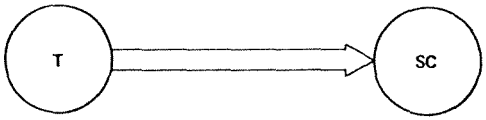


FIG. 9.8. Cause.

The tendency T is forcing the particular component to change state. For example, if the magnetic field starts the tendency of pulling, the switch will change state from closed to open (SC).

The third type of link is **propagate:**

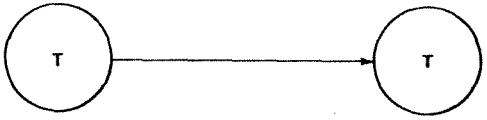


FIG. 9.9. Propagate.

The laws for connections derive T2 from T1, thus propagating values through the device topology. For example, if the current through the switch is zero, the current through the battery also becomes zero.

The final type of link is **antagonism:**

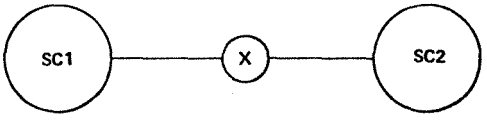


FIG. 9.10. Antagonism.

Both state changes SC1 and SC2 cannot hold simultaneously and are therefore termed antagonistic. For example, the switch state open is antagonistic to the switch state closed.

A representation for the functioning of the buzzer represented in this epistemology is shown in Fig. 9.11.

This representation is easily constructed from Fig. 9.6 that was obtained from the transition table. The topology of Fig. 9.11 is isomorphic to that of Fig. 9.6, except for the addition of antagonism links. The edges of Fig. 9.6 are represented by the appropriate functional link type. Every edge that represents an attribute value being successfully tested for a state transition is represented by a cause link. Every edge where a definitional rule is used is represented by an enablement link. Every edge where a connection law is used to derive a new attribute value from an old one is represented by a propagate link. Note that this representation for function cannot be constructed from the transition table alone. To identify the origin of the changes in the table and their respective type requires referring back to the particular component rules that produced the change.

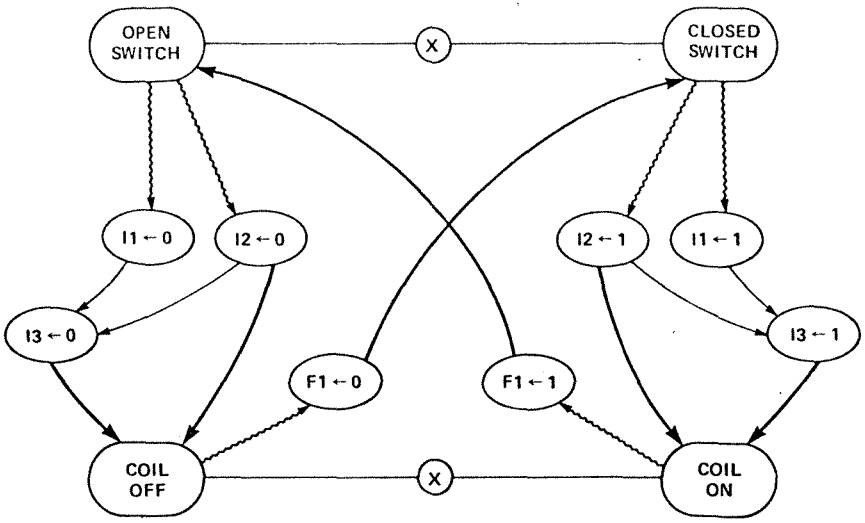


FIG. 9.11. Causal attribution of buzzer.

In order to determine the behavior of the buzzer over time, one can just step through the links in Fig. 9.11 reading the sequence of values at successive time increments.

CAUSALITY PRINCIPLES AND ENVISIONMENT

The representation of function meets one of the desiderata we were originally seeking for mental models but could not get: The next state (necessarily unique) of the composite device is directly and solely determined by the tests of component rules on the previous composite state. This desiderata subsumes the weak causality principle, which is important because it ensures that the simulation can be done efficiently and that every component state change be caused by the previous composite state. Because the causes for these state changes are identifiable, they can be used in identifying reasons for faulty and correct predictions, and thus the principle plays a fundamental role in learning.

This desiderata is partially met in the models, because state transitions can only be made by consequences of component rules, but this restriction only meets the weak causality principle. It does not guarantee the strong causality principle of unique successor state is met, because for example, a composite state may have multiple successor composite states. Furthermore, a great deal of deduction may be required to determine which states to eliminate. Thus, we see the simulation process (P1) responsible for constructing envisionments in clear perspective: Because its result does not necessarily obey the strong causality

principle, a subsequent analysis (P2) is required to construct a representation that does obey the strong causality principle.[6]

A CRITIQUE OF THE RIEGER AND GRINBERG'S APPROACH

The goal of Rieger and Grinberg's research is to study how humans might represent cause-effect knowledge about physical mechanisms. We do not question their epistemology, nor that their representation for function is valuable, but rather we question whether a representation of function alone plays a major role in understanding physical mechanisms. Although their representation is intuitively appealing, it is mainly so because of its similarity to superficial characteristics of human explanation. They did not consider how the representation of structure could be constructed from the device itself.

What distinguishes our position from Rieger and Grinberg's on a theory of mechanism is our concern for a representation of the physical device itself, and more importantly, a theory of mechanism that deals with the relationship between the physical device (its structure) and the overall causal behavior manifested by that device (its function). A representation for either the structure or the function alone is by itself insufficient. No matter how each is represented, the important issue is the relationship between the two.

A theory of mechanism that is only concerned with functionality cannot address many of the questions we expect such a theory to answer. For example, the simple question, "What happens if the clapper switch is removed?" cannot be answered at all, because that question is asking what change in functionality results if the structure is altered. For their theory to have utility, Rieger and Grinberg have had to confuse their epistemology of functionality with some primitives that are inherently structural. For example, in their representation of the buzzer, two states "Open Switch" and "Closed Switch" need to be related by a link that states that both cannot hold true simultaneously. This link, like every other in their theory, needs to be added explicitly. If their theory included an adequate notion of structure, this would be unnecessary, because both of these states refer to the same clapper switch, and a single part can only be in one state at one time.

Although Rieger and Grinberg's notion of simulation may seem similar to ours, their's is used for an entirely different purpose. Their representation of the

[6]The buzzer is really too simple to illustrate the utility and difficulty of the analysis process P2. An examination of the component models reveals that the composite buzzer model will obey the strong causality principle completely. This can be seen by the fact that the topology of Fig. 9.11 is isomorphic to the topology of the rules. However, consider the case where the buzzer had two clappers. Here, there is some ambiguity as to which opens first, and this must be resolved in the representation of function.

functionality is specifically designed to apply for various inputs. What they call simulation is a reasoning process within that representation to determine the response of the device, given a particular stimulus. Stimuli not included in the original representation cannot be dealt with. Our use of simulation is as a step in a reasoning process on the representation of the structure to determine its functionality. This representation of structure is derived by a straightforward process utilizing a set of models valid for all devices in some class. As a consequence, Rieger and Grinberg's representation cannot be used to simulate devices that are even minutely different in structure, whereas ours can be used for any of a broad class of devices.

Another way to examine the distinction between the two theories is to ask what questions each can answer about a device. Rieger and Grinberg's representation can answer questions about the behavior resulting from some external stimulus, or about the consequences of removing or changing a causal link. However, it cannot answer questions about the consequences of some component being removed or faulted, or about behavior at some extreme section of its operating range. It is not surprising that the ability to answer these questions is necessary for designing or troubleshooting. What is more important is that it is necessary for adequately coping with malfunctions in devices. Having such a robust understanding of the buzzer requires the ability to determine the change in function that results from a change in structure.

DELETION PRINCIPLE

Our model for the buzzer still suffers a number of difficulties, and the remaining sections of the chapter sketch out some of these problems and what can be done about them.

Although the model conforms to the locality principles, it still has some of its function embedded in its structure. The model violates the deletion principle: models should not predict that a machine still functions when a vital part is removed. Application of this principle checks whether the functionality of some part is implicitly embedded in the structure of other parts. The models for the buzzer predict that it will continue to function even when the battery is replaced by a short circuit, because the model for a clapper switch assumes it is connected to a battery. Removing this assumption gives the following model for the clapper switch:

```
Information terminals: i1, i2, f1
SWITCH : OPEN:
              i1 <- 0, i2 <- 0
              if f1 = 0, switch will become CLOSED.
         CLOSED:
              i1 <-> i2
              if f1 = 1, switch will become OPEN.
```

Unfortunately, the models now make no predictions at all, because after the switch changes from state open to closed, no rules derive a value for the currents, and thus the coil fails to change state from off to on. The problem here is that the models used for the connecting wires fail to refer to voltage. Using the hydraulic metaphor, current is the amount of material that is flowing, and voltage is the pressure on that material causing it to flow. The battery is the supplier of the necessary voltage in the buzzer. Because the models take neither voltages nor the battery, into account, it is not surprising that they are inadequate to characterize the behavior of the buzzer.

DISCONTINUITY

Another problem with the models is that they fail to characterize what actually occurs during the state transition. As the coil changes state from on to off, the current through the coil must change from 1 to 0. For a current to change from 1 to 0, it must be .5 sometime during the change. It could conceivably change from 1 to .5 to 10 to .5 to -100 to 0. A physical quantity cannot change abruptly and must vary continuously (i.e., if we make the time increments small enough, the changes in current must become arbitrarily small). However, as far as these models are concerned, the current changes discontinuously from 0 to 1. Because the models fail to characterize what happens during a state transition, they cannot completely describe how a transition is caused. It is incorrect to say that a transition can be caused by some attribute being at a particular value. For example, the rule

if I = 0, then coil will become OFF

is false. If no current is flowing through the coil, it is not the case that the coil will *become off,* rather it must have been off for current to be absent. This is not a situation where mythical or real time is at issue. The previous models used mythical time to model the transition, but that is an incorrect analysis of the actual physics. A transition is not caused by an attribute having some value but by some attribute changing value. Furthermore, the state transition is caused when the value crosses some threshold. For example, in the case of the coil, if the current drops from 1 to something slightly less than 1, the transition will probably not occur. But certainly, when the current almost reaches 0, the coil must have changed state. This may suggest that a coil should have three states: on, off, and between, but the transitions would still be discontinuous.

THE ROLE OF MOMENTUM IN QUALITATIVE REASONING

In order to model the behavior during transitions, the models will have to distinguish between the value of an attribute and the changing of an attribute. Space

precludes us from presenting the set of resultant models, but some surprising consequences arise. In order to maintain internal consistency, we have to consider the notion of momentum. This should not be too surprising, because the buzzer could not possibly function if it did not have momentum, and each of the models discussed so far has implicitly assumed it. Consider what happens just as the magnetic field becomes strong enough to begin to lift up the clapper switch. The circuit will begin to open, less current will flow, and the magnetic field will begin to decrease. Because the magnetic field was just strong enough to begin to lift the clapper switch, this decrease will weaken the field enough so that it can lift the clapper switch no further. Therefore, the clapper switch will remain stuck at a position that just begins to open the circuit and the buzzer will fail to function.

The real buzzer functions because the attraction of the magnetic field acts on the mass of the switch, accelerating it. By the time the switch begins to open, this acceleration has built up a velocity in the switch. Although when the field dies, the acceleration due to gravity (or a restoring tendency in the clapper arm) pulls the switch back to its resting position, it will take some time before this acceleration even cancels out the velocity built up in the switch by the momentary magnetic field. Thus, even after the field dies, the switch will continue to open. The buzzer buzzes but only because the switch possesses momentum. In the original diagram Fig. 9.3, time increment t = 2 actually described a very complex event of the switch continuing to move toward the magnet, momentarily holding still and then beginning to fall back to its resting position. Having a distinction between the value of an attribute and a change in an attribute enables the new models to deal explicitly with momentum, because the essential point of momentum is that forces often act on changes in attribute values (their derivatives), not on the values themselves. This is especially important, because derivatives are usually only associated with quantitative or analytic analysis, and thus are not often thought to be relevant to the common sense understanding of mechanistic devices.

REFERENCES

Brown, J. S., & Burton, R. Multiple Representations of Knowledge for Tutorial Reasoning. In D. Bobrow & A. Collins (Eds.), *Representation and understanding*. New York: Academic Press, 1975.

Brown, J. S., Collins, A., & Harris, G. *Artificial intelligence and learning strategies* (B. B. N. Report 3634, I.C.A.I. Report 6. Cambridge: B. B. N., 1977. (Also in *Learning Strategies*. New York: Academic Press, 1978.)

de Kleer, J. "Causal and Teleological Reasoning in Circuit Recognition. *Artificial Intelligence Laboratory* (TR-529). Cambridge: M. I. T., 1979.

Rieger, C., & Grinberg, M. The Declarative Representation and Procedural Simulation of Causality in Physical Mechanisms. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*. 250-255, 1977.