REASONING ABOUT HIDDEN MECHANISMS

Richard J. Doyle
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, California 91109

Abstract

I describe an approach to the problem of forming hypotheses about hidden mechanisms within devices --- the "black box" problem for physical systems. The approach involves enumerating a set of physical and causal constraints and enumerating difference causal structures for devices, placing an ordering on these hypothesis types, and carefully controlling the generation of hypotheses. I relate in detail the performance of an implemented causal modeling system on the surprisingly puzzling pocket tire gauge. Results from several examples indicate that the ideas presented support capabilities for maintaining manageably sized hypothesis sets and for making fine distinctions among hypotheses.

1. Figuring Out How Things Work

The process of constructing and refining physical models to account for observations is an important form of reasoning. In this paper, I investigate the modeling process itself. The domain is mechanical, electrical, and thermal devices --- designed physical systems. The research goal is to articulate a set of principles which support capabilities for hypothesizing manageably small sets of physically plausible device models, and for making fine distinctions among those models. I have developed a modeling system --- called JACK --- which addresses these questions and produces abstract causal models of several physical systems, including a toaster, a pocket tire gauge, a bicycle drive, a refrigerator, and a home heating system.

The importance of the modeling problem arises from its ubiquity. The need to understand how things work inevitably arises in the course of other problem solving tasks. In the physical system domain these tasks include diagnosis, monitoring, and design.

My approach to making the modeling problem tractable in the physical system domain has two thrusts. One thrust involves applying a set of constraints which embody physical and causal principles to prune hypotheses. The other thrust involves enumerating different forms for hypotheses, placing an ordering on these forms, and using this ordering to carefully control the generation of hypotheses.

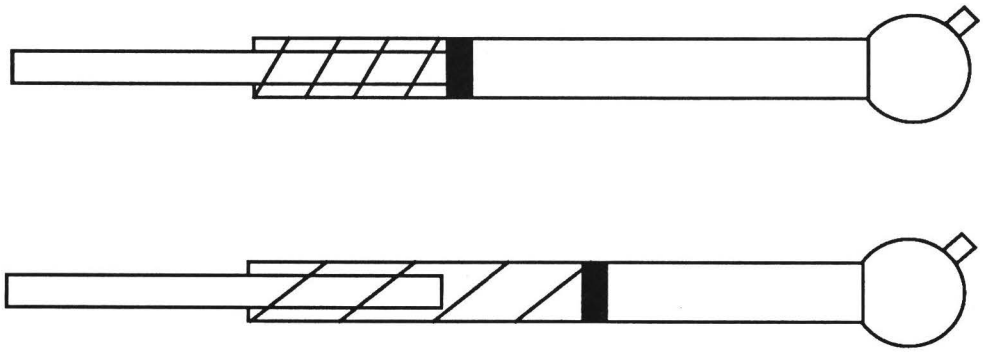

Figure 1. Why does the slide remain stationary?

2. A Scenario

The pocket tire gauge is an excellent example of a device for which the modeling problem is surprisingly thorny. Its range of behavior is quite small, yet this behavior is baffling.

For example, if the motion of the slide in a tire gauge is a response to air pressure, why doesn't the slide slam all the way to the end of the cylinder? One possible explanation involves an equilibrium state within the cylinder. There may be an opposing force --- due to a spring, for example --- which balances the air pressure. However, why doesn't the slide slip back into the cylinder when the gauge is removed from the tire? The conjectured spring force then should be the only active one. See Figure 1.

3. The Causal Modeling Task

The task of the causal modeling system JACK is to conjecture configurations of mechanisms inside the "black box" which are consistent with the externally observable behavior of a device.

There are two inputs to the causal modeling system: one is a description of the externally observable behavior of a device; the other is a set of mechanisms. The output is a set of compositions of those mechanisms, each explaining the behavior of the device.

The causal modeling problem can be stated as a graph problem. The nodes of the graph correspond to the events of a device --- changes in the values of its quantities. The arcs of the graph correspond to the mechanisms which map events to other events.

The task is to construct a set of directed graphs consisting of mechanisms and intermediate events which connect known input events to known output events. See Figure 2. These *causal graphs* are the output of the causal modeling system.

A set of observable events forms the periphery of each graph to be constructed. The mechanisms and intermediate events correspond to hypotheses about what hidden mechanisms may exist and what unobservable events may take place inside the black box.

4. The Approach

In this section, I describe briefly the set of constraints and the ordering of hypothesis types in the physical system domain which are at the crux of my approach to the causal modeling problem.

4.1 Physical and Causal Constraints

The constraints I enumerate here concern how different observable aspects of the behavior and structure of physical systems are conserved or transformed across mechanisms. All hypotheses about mechanisms within devices are subject to these constraints.

The *type* constraint concerns the types of quantities in a physical system. A mechanical coupling is an admissible explanation for a cause whose type is rate of position and an effect whose type also is rate of position.

The *delay* constraint concerns the times of occurrence of events in a physical system. Electricity or a rigid coupling, whose propagation times are essentially instantaneous, are consistent hypotheses for a cause and effect which are perceptually simultaneous.

The *sign* constraint concerns the signs of the values of quantities in a physical system. Flow in a closed system implies a decrease in amount at the cause and an increase at the effect, or vice versa.

The *direction* constraint concerns the orientations in space of quantities in a physical system. A spring, which produces a reversal in the direction of motion, is a consistent explanation for a motion followed by a motion in the opposite direction.

The *magnitude* constraint concerns the magnitudes of the values of quantities in a physical system. A rigid coupling, which transfers motion with no loss, can be a causal explanation only for motions of the same magnitude.

The *alignment* constraint concerns the relative values of quantities in a physical system. For a non-rigid coupling such as a string, the position of the cause must be greater than the position of the effect, along the direction of motion.

The *bias* constraint concerns the directions of change of quantities in a physical system. A ratchet allows motion in one direction but not in the opposite direction.

The *displacement* constraint concerns the locations of objects in a physical system. Thermal expansion cannot account for a temperature change in one physical object

and a motion in another because thermal expansion takes place entirely within one physical object.

The *medium* constraint concerns the connections between objects in a physical system. For example, gas flow is an admissible hypothesis when two physical objects are joined, but is untenable when they are separated.

4.2 An Ordering on Hypotheses

The simplest type of causal graph involves only linear mechanism paths between input events and output events. However, linear mechanism paths may be extended into branching mechanism interactions. Three types of mechanism interaction are distinguished: *enablement*, where one mechanism arranges for the preconditions of another mechanism to become satisfied; *disablement*, where one mechanism arranges for the preconditions of another mechanism to become unsatisfied; and *equilibrium*, where the contributions of separate mechanisms come into balance.

An example of enablement is a switch being closed and permitting the flow of electricity. An example of disablement is a latch being engaged and arresting a motion. An example of equilibrium is the steady level of water in a sink when the flow in at the faucet balances the flow out at the drain.

The causal modeling system does not extend all linear mechanism hypotheses. In the interest of keeping the hypothesis set manageably small at all times, a set of heuristics is employed for deciding when to consider hypotheses involving mechanism interactions. These heuristics capture manifestations of the following principle: *Incomplete hypotheses often exhibit characteristic deficiencies.* These signatures indicate into what form of interaction hypothesis a deficient linear hypothesis should be extended.

4.3 Heuristics for Recognizing Interactions

Enablements are characterized by unexplained delays. Once a pending mechanism becomes enabled however, the resulting effect is always as expected. The exception is a possible decrease in magnitude as in the case of say, a half-open valve. The heuristic for recognizing enablement situations is:

> Either exactly the delay constraint is violated or
> exactly the delay and
> magnitude constraints are violated.

The signature for disablements is an unexpected zero value occurring after a non-zero effect is expected. The heuristic for recognizing disablement situations is:

> Exactly the delay, sign, magnitude, and
> bias constraints are violated
> and the value of the effect is zero
> and the effect is not at a limiting value.

Equilibria also are characterized by an unexpected zero value when the expected effect is non-zero. The zero value may occur after the expected time of occurrence of a non-zero effect. The heuristic for recognizing equilibrium situations is:

> Either exactly the sign, magnitude, and
> bias constraints are violated
> or exactly the delay, sign, magnitude, and
> bias constraints are violated
> and the value of the effect is zero
> and the effect is not at a limiting value.

4.4 Mechanisms

Mechanisms are the building blocks for forming causal explanations. There are 50 mechanisms in the vocabulary of the program JACK, including mechanical linkages, electricity, heat flows, thermal expansion, evaporation and condensation, gravity, switches, latches, valves, etc. Each mechanism is defined in terms of the constraints.

Every mechanism has a specific quantity type associated with its cause and with its effect. The time constant of a mechanism determines the range of delays it can account for. The sign of the quantity dependence associated with a mechanism restricts the sign conservations or transformations it can explain. The deflection associated with a mechanism determines the changes of direction it can account for. The efficiency of a mechanism determines what changes in magnitude it can explain. The alignment relation associated with a mechanism places a restriction on the relative values at cause and effect. The bias relation of a mechanism constrains the directions of change at cause and effect. The distance associated with a mechanism determines the displacements between cause and effect it can account for. The medium associated with a mechanism indicates the structural relation which must obtain between cause and effect.

An example of a mechanism definition appears below:

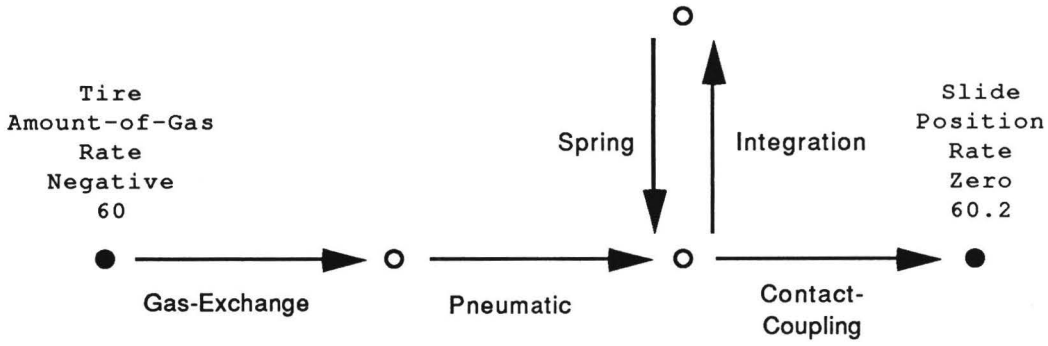

Figure 2. A tire gauge hypothesis.

(DEFMECHANISM *Contact-Coupling*
:cause-type {TYPE *Position Rate*}
:effect-type {TYPE *Position Rate*}
:time-constant 2^{-3} : ++
:sign *Positive*
:deflection *Parallel*
:efficiency 2^{0} : 2^{0}
:alignment *Less*
:bias {*Up-Up Down-Down*}
:distance *Different*
:medium *Touches*)

4.5 Events

Device events also are represented in terms of the constraints. The values propagated along a mechanism path for each of these constraints make up a detailed description of the events which are expected to take place along the path. For example, an event node describing the motion of a proposed hidden object within the tire gauge is:

{EVENT
:type {TYPE *Position Rate*}
:delay 2^{-3} : 2^{7}
:sign {*Positive*}
:direction {*Parallel Opposite Perpendicular Skewed*}
:magnitude 2^{-10} : 2^{3}
:alignment {*Less Equal Greater*}
:bias *Positive*
:displacement {*Same Different*}
:medium {*?physical-object-2*}}

4.6 Propagation and Combination Rules

Each proposed causal model of a device is simulated by propagating and combining values for the physical and causal constraints along the proposed mechanism paths. Predicted values describe expected events which must be compatible with observed events for a hypothesis to be admitted.

Some examples of propagation rules follow. A non-zero sign {*Negative Positive*}, when propagated across a mechanism with a bias towards increase in the effect: {*Down-Up Up-Up*}, becomes unambiguous: {*Positive*}. Delays are propagated across a mechanism by adding the time lag associated with the mechanism. This time lag is computed by multiplying the distance across the mechanism by the time constant associated with the mechanism. For each device, a default distance is established. The set of physical objects {*Tire*}, when propagated across a mechanism whose medium is *Joined-To*, is the set of physical objects {*Cylinder*}, providing the relation {*Tire Joined-To Cylinder*} has been asserted to be true, and no other relations {*Tire Joined-To* *} have been asserted to be true.

The contributions of interacting mechanisms are combined at points of interaction. For example, in enablement/disablement situations, delay is measured from the *later* of the interacting causes. In other words, no effect occurs until all causes are in place. The handling of delay in equilibrium situations is different. An intermediate effect may occur at the time of the earliest contribution; however, the effect which is the result of interaction occurs at the time of the latest contribution.

5. Reasoning About the Tire Gauge

In this section, I work through a detailed example of hypothesis construction for the tire gauge. The hypothesis is shown in Figure 2.

The device event {*Tire Amount-of-Gas Negative 60*} is taken to be the cause and the device event {*Slide Position Rate Zero 60.2*} is taken to be the effect. One of the generated hypotheses is the linear mechanism path {*Gas-Exchange Pneumatic Contact-Coupling*}. The seed event node computed from the cause event is:

{EVENT
:type {TYPE *Amount-of-Gas Rate*}
:delay *0:0*
:sign *Negative*
:direction *Parallel*
:magnitude $2^0 : 2^0$
:alignment {*Less Equal Greater*}
:bias *Negative*
:displacement *Same*
:medium *Tire*}

The target event node computed from the effect event is:

{EVENT
:type {TYPE *Position Rate*}
:delay $2^{-2} : 2^{-2}$
:sign *Zero*
:direction *Skewed*
:magnitude *0:0*
:alignment {*Less Equal Greater*}
:bias *Zero*
:displacement *Different*
:medium *Slide*}

The event node which represents the effect of the *Gas-Exchange*, *Pneumatic*, and *Contact-Coupling* mechanisms is computed via the propagation rules for the constraints. This event node is:

{EVENT
:type {TYPE *Position Rate*}
:delay $2^{-10} : 2^0$
:sign {*Positive*}
:direction {*Parallel Opposite Perpendicular Skewed*}
:magnitude $2^{-7} : 2^0$
:alignment *Less*
:bias {*Positive*}
:displacement {*Same Different*}
:medium {*?physical-object-3*}}

This event node is incompatible with the target event node. In particular, the sign, magnitude, and bias constraints are unsatisfied. However, this partial failure triggers the equilibrium interaction heuristic.

One of the proposed equilibrium hypotheses involves the mechanism path {*Integration Spring*} which splits from and rejoins the given mechanism path just before the *Contact-Coupling* mechanism.

The event node which represents the contribution of the *Gas-Exchange* and *Pneumatic* mechanisms before the split is:

{EVENT
:type {TYPE *Position Rate*}
:delay $2^{-10} : 2^0$
:sign {*Positive*}
:direction {*Parallel Opposite Perpendicular Skewed*}
:magnitude $2^{-7} : 2^0$
:alignment {*Less Equal Greater*}
:bias {*Positive*}
:displacement {*Same Different*}
:medium {*?physical-object-2*}}

The event node which represents the contribution of the *Integration* and *Spring* mechanisms before the rejoin is:

{EVENT
:type {TYPE *Position Rate*}
:delay $2^{-3} : 2^3$
:sign {*Negative Zero*}
:direction {*Parallel Opposite Perpendicular Skewed*}
:magnitude $0 : 2^3$
:alignment {*Less Equal Greater*}
:bias {*Negative*}
:displacement {*Same Different*}
:medium {*?physical-object-2*}}

The event node which represents the additive combination of these two contributions is:

{EVENT
:type {TYPE *Position Rate*}
:delay $2^{-3} : 2^3$
:sign {*Negative Zero Positive*}
:direction {*Parallel Opposite Perpendicular Skewed*}
:magnitude $0: 2^3$
:alignment {*Less Equal Greater*}
:bias {*Negative Zero Positive*}
:displacement {*Same Different*}
:medium {*?physical-object-2*}}

Finally, the event node which represents propagation through the *Contact-Coupling* mechanism after the equilibrium interaction is:

{EVENT
:type {TYPE *Position Rate*}
:delay $2^{-3} : 2^{3}$
:sign {*Negative Zero Positive*}
:direction {*Parallel Opposite Perpendicular Skewed*}
:magnitude $0: 2^{3}$
:alignment *Less*
:bias {*Negative Zero Positive*}
:displacement {*Same Different*}
:medium {*?physical-object-3*}

This event node is compatible with the target event node. This hypothesis is admitted.

The program JACK generates a number of additional hypotheses for the tire gauge. Among these is a disablement interaction hypothesis to explain the halting of the slide's motion. This model for the tire gauge also involves pneumatic motion of a hidden physical object. However, in this case the motion of the hidden object displaces not a spring but a valve. When the valve is closed, the flow of gas is disabled, and the motion of the slide --- transmitted along a mechanical coupling from the hidden object --- also stops. Thus an impulse of displaced gas is responsible for the start-and-stop motion of the slide.

In another proposed model, an equilibrium hypothesis, there are two pathways for gas flow. One pathway is short and generates pneumatic motion of the slide. The other pathway is longer and is directed backward to oppose the flow along the first pathway. When the second gas flow collides with the first an equilibrium state is created and motion of the slide stops.

6. Empirical Results

The program JACK constructs causal graphs which connect observable events of a device. Causal graphs which connect the same subset of observable events are collected into "grey compartments". Grey compartments form a useful abstraction space from which to reason about a device. They answer the question "Which events affect one another?" rather than "How do events affect one another?". Grey compartments are decoupled because they intersect at observable events. Complete and consistent models of a device can be built by chaining together the causal explanation fragments represented by grey compartments, starting at the known input events of a device and ending at the known output events. Within each grey compartment there may be several different mechanism configurations which explain the same behavior.

Table 1 shows the number of grey compartments and causal graphs within those grey compartments admitted by the program JACK for several implemented device examples. l_{max} is the length of the longest mechanism path in any causal graph for the given device. p_{max} is the greatest number of interacting paths in any causal graph for the given device. The number of causal graphs is reported as the sum of the causal graphs in the individual grey compartments to emphasize that hypotheses in different grey compartments are mutually independent.

In a set of experiments in which the physical and causal constraints were utilized in isolation from one another, the *type* constraint was found to be the single most effective source of pruning power, followed by the *delay* constraint. The pruning ratio associated with the mechanism interaction recognition rules was approximately 150; in other words, roughly one out of every 150 linear mechanism path hypotheses was extended into a mechanism interaction hypothesis.

7. Relation to Other Work

Several approaches to causal and qualitative reasoning have appeared in the literature. Seminal works among

Device	l_{max}	p_{max}	Grey Compartments	Causal Graphs
Toaster	2	2	12	93
Tire Gauge	4	2	2	400
Bicycle Drive	2	2	3	31
Refrigerator	4	2	2	222
Home Heating	3	3	8	464

Table 1. Number of hypotheses admitted.

these include Forbus' Qualitative Process Theory [Forbus 85], de Kleer and Brown's qualitative physics based on confluences [de Kleer and Brown 85], and Kuipers' method for qualitative simulation [Kuipers 86]. One of the lessons learned from these efforts is that causal and qualitative reasoning subsumes several complementary forms of inference.

In my work, the several constraints serve as multiple representations, supporting reasoning about different observable dimensions of the behavior and structure of physical systems. Collectively, these constraints support reasoning about dynamics --- which changes occur?, time --- when do events occur?, physical objects --- where do events occur?, topology, what are the causal pathways?, thresholds --- what new values are reached?, and preconditions --- which mechanisms are active and which are inactive?

Shrager, in his research on instructionless learning [Shrager 87], also investigates the modeling problem. He focuses on a cognitive model of device hypothesis construction in humans while my emphasis is on the sources of constraint which make the problem tractable.

8. Conclusions

There are a number of assumptions and limitations inherent in the approach to modeling I have described in this paper. Firstly, there are closed-world assumptions, both at the level of mechanisms and at the level of causal graph structures. In particular, mechanisms such as pulleys, friction, and magnetism, to name a few, are not described, and causal structures such as iterative cycles, devices with state, and certain couplings between mechanisms (such as a fluid flow supporting a heat flow) are not described.

Moreover, second and higher order derivatives are not represented, and dependencies between quantities are assumed to be linear and monotonic. The representations for physical structure are fairly impoverished. Finally, there is limited ability to reason in the teleological domain.

Nonetheless, I have addressed the problem of how to constrain the formation of hypotheses about mechanisms within physical systems. I have enumerated a set of constraints based on physical and causal principles which support reasoning about several observable aspects of devices. I have enumerated a set of causal structures for devices. I have dealt with the complexity vs. completeness problem by placing an ordering on these hypothesis types and designing a set of heuristics for recognizing when failed hypotheses should be extended into more complex hypotheses. These rules are based on the principle that incomplete hypotheses often exhibit characteristic deficiencies. Results from several implemented examples indicate that these ideas support capabilities for maintaining manageably sized hypothesis sets and for making fine distinctions among hypotheses.

Acknowledgements

In performing the work described in this paper, I have benefited from discussions with Jonathan Amsterdam, Randall Davis, Tomas Lozano-Perez, Karl Ulrich, and Patrick Winston.

This report describes work done while the author was at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for this laboratory's Artificial Intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124.

References

[de Kleer and Brown 85] Johan de Kleer and John S. Brown, "A Qualitative Physics Based on Confluences," in *Qualitative Reasoning About Physical Systems*, D. Bobrow, ed., MIT Press, 1985.

[Doyle 88] Richard J. Doyle, "Hypothesizing Device Mechanisms: Opening Up the Black Box," Report TR-1047, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1988.

[Forbus 85] Kenneth D. Forbus, "Qualitative Process Theory," in *Qualitative Reasoning About Physical Systems*, D. Bobrow, ed., MIT Press, 1985.

[Kuipers 86] Benjamin J. Kuipers, "Qualitative Simulation," *Artificial Intelligence*, **29**, 1986.

[Shrager 87] Jeff Shrager, "Theory Change via View Application in Instructionless Learning," *Machine Learning*, **2**, 1987.

"How Things Work," **1-4**, Edito-Service S. A., Geneva.