CREATING CAUSAL MODELS

Richard J. Doyle
Massachusetts Institute of Technology

Extended Abstract

Submitted to the Organizing Committee
of the
Workshop on Qualitative Physics

February 9, 1987

The Problem

The task of developing qualitative and causal reasoning systems to perform problem solving on physical systems has two aspects: (1) Designing representations for structure, behavior, and causality within which to describe the physical systems of interest and their constituent objects and processes, and (2) Developing algorithms which operate on the chosen representation to efficiently draw inferences to perform the given problem solving task – simulation, troubleshooting, etc. Embedded in this methodology is an assumption that complete models for the physical systems of interest are known and can be rendered in the chosen representation.

My research concerns the problem which arises when this assumption is removed: How to *create* the models of physical systems on which problem solving can be performed. A more specific statement of my research problem is: Given (1) a description of the *externally observable* structure and behavior of a physical system, and (2) a vocabulary of primitive mechanism schemata which provide causal explanations for specific kinds of behavior, the task is to hypothesize configurations of mechanisms which provide a complete causal account of the observable behavior of the physical system.

This *causal modelling* problem can be cast as theory formation – the physical system being modelled is considered static and the task is to construct and incrementally refine a description of the device which explains an ever-growing set of examples of the behavior of the device. Furthermore, each hypothesis is a proposed design – the observable behavior serves as a set of functional specifications and the task is to construct a configuration of mechanisms from a given vocabulary which realizes those specifications. Clearly, there are many possible designs for a given device and some means of constraining hypotheses must be employed in order to make the problem tractable. This issue is the main focus of my work.

Motivation

An effective causal modelling system must entertain only reasonable hypotheses about unobservable configuations of mechanisms. One goal of my research, from an academic viewpoint, is to enumerate the kinds of constraints which designs implicitly satisfy. Knowledge of these constraints will enable my causal modelling system to separate reasonable from unreasonable hypotheses. The identification of such knowledge should prove useful to other research endeavors which address reasoning about the behavior of physical systems.

From a practical viewpoint, the causal modelling problem may not seem all that important or realistic at first glance. Are not schematics and other documentation generally available for physical systems intended for actual installation and operation? However, quite often such documentation is hopelessly obsolete, contains unfortunate abstractions, etc., such that a direct translation into a representation appropriate for a qualitative reasoning system is not feasible. In these cases, an automatic modelling capability would prove useful.

The Domain

I will evaluate the performance of my causal modelling system on devices which contain electrical, mechanical, and thermal mechanisms. The order of complexity I have in mind is that found in the

typical household gadget. Examples are a pocket tire gauge, a toaster, a bicycle drive with coaster brake, a door knob with key and lock, a door closer, a camera, and an automobile turn signal.

The Approach

The most difficult aspect of the causal modelling problem is constraining the generation of hypotheses about hidden configurations of mechanisms within a device. The approach I am taking proceeds from the observation that the devices whose mechanism structures are to be inferred are *designed* artifacts. All designs implicitly reflect a host of constraints which fall into two categories: (1) general inviolable constraints due to physics and causality and (2) specific pragmatic constraints concerning cost, available components, size, weight, etc. These constraints collectively serve as my definition of reasonableness for hypotheses about mechanism configurations. My approach to making the causal modelling problem tractable involves enumerating and utilizing knowledge of these design constraints to generate manageable sets of hypotheses.

At this stage of my investigation, I am concentrating on enumerating the always relevant physical and causal design constraints. Ultimately, I may need to introduce knowledge of pragmatic constraints as well. The physical and causal constraints I have identified are:

- *Energy Type* – Mechanisms have well-defined input and output energy types. For example, thermal expansion transforms thermal energy into motion. Couplings transfer motion, possibly changing its type (e.g. linear to rotary). Hypothesized mechanism configurations must introduce any necessary energy transformations from initial causes to final effects.
- *Spatial Proximity* – Some mechanisms manifest causation from one site to another (electricity); others concern processes which take place within a single physical object (photochemical). Those mechanisms which involve spatially disparate causes and effects in turn fall into two categories: Some have an explicit medium (couplings, material flow), whereas fields (gravity) and radiation (light transmission) propagate freely without an observable medium (so-called action at a distance). Hypothesized mechanism configurations for spatially distinct initial causes and final effects must include at least one mechanism to account for the site change.
- *Temporal Proximity* – Some mechanisms involve a delay between manifestation of cause and effect (conductive heat) while others propagate causation instantaneously (electricity). Hypothesized mechanism configurations for temporally disjoint initial causes and final effects must incorporate at least one mechanism which introduces a delay.

There is an important alternate explanation for temporal delays between cause and effect. Mechanisms may have *thresholds* among their preconditions (switches, latches) which introduce arbitrary delays between the moment the primary cause of a mechanism is established and the moment its entire set of preconditions is satisified.

- *Efficiency* – Mechanisms have an inherent efficiency from which governs the magntitude of the effect, given the magnitude of the cause. For example, hydraulic systems are generally of high efficiency while thermal systems are typically of lower efficiency. Factors such as the length and opacity of a medium contribute to efficiency. Hypothesized mechanism configurations must have a composite efficiency sufficient to account for the magnitude of the final effect.

- *Magnitude* – Mechanisms may also have an inherent maximum magnitude, regardless of the efficiency of the configuration of mechanisms which delivers the cause. For example, thermal expansion never results in large motions. Hypothesized mechanism configurations must not include mechanisms which impose a maximum magnitude smaller than the magnitude of the final effect.
- *Direction* – Those mechanisms which involve forces and motions (both linear and rotary) have well-defined directions (or senses, in the cases of torque and rotary motion). For example, compression springs provide a force in the direction opposite from the direction leading from the compressing object to the spring; gravitationally-induced motion is always downward. Hypothesized mechanism configurations which involve forces and motions must introduce any necessary deflections from initial causes to final effects.

The constraints above are used to filter mechanism configuration hypotheses by determining if the expected behavior of a device under each hypothesis is consistent with the actual observed behavior of the device.

- *Misbehavior* – Misbehavior in a device also can be employed to filter hypotheses. Misbehavior is simply more behavior to reason about and account for. Knowledge of what constitutes misbehavior is typically independent of knowledge of the mechanisms within a device. For example, overexposed film and burnt toast is recognizable without working knowledge of cameras and toasters. If failure modes for mechanisms are known, mechanism configuration hypotheses can be tested further by determining if there is a way for the proposed mechanism configuration to fail which is consistent with the observed misbehavior.

This is troubleshooting starting with an incomplete model of the working device. Theory formation and troubleshooting proceed hand-in-hand, each constraining the other, in a kind of hand-shaking back-and-forth propagation.

The Vocabulary of Mechanisms

The vocabulary provided to my causal modelling system contains broad categories of mechanisms.

- *Propagations* are mechanisms which causally connect events at spatially separate sites. No energy transformation is involved. Propagations are either channelled (with an explicit medium) or free. The channelled propagations include all couplings (including contacts, fasteners, spindles, gears), material flow, conductive heat, and electricity. The free propagations are radiative heat and light transmission.
- *Transformations* are instantaneously manifesting mechanisms within a single physical object which involve energy transformations. This class includes the electrothermal, electrophotic, electromechanical, expansion, friction, thermochemical, and photochemical mechanisms.
- *Forces* are also channelled or free. The channelled forces include springs (including compression springs, tension springs, torsion springs) and pressure (including pneumatic and hydraulic). The free forces are fields: gravity and magnetism.
- *Thresholds* are mechanisms whose effects appear in inequalities among the sets of preconditions of other mechanisms. Achieved thresholds can render other mechanisms active or inactive.

Examples of thresholds are switches, latches, ratchets, and valves (including slide valves, stem valves, and irises).

Unfortunately, space restrictions prevent me from providing a detailed description of a mechanism schema.

This finite vocabulary of mechanisms clearly embodies a closed-world assumption. The same comment applies to the set of failure types associated with each mechanism for reasoning about misbehavior.

Examples

The device which is the current focus for my implementation efforts is a toaster. Figure 1 shows the external observation of a toaster which is input to my causal modelling system.

```
;;;TIMELINE;;;
(start) ;t=0
(assert-quantity LEVER-POSITION Amount Up Rate Zero)
(assert-quantity DIAL-ANGLE Amount LM Rate Zero)
(assert-quantity CARRIAGE-POSITION Amount Up Rate Zero)
(assert-quantity COILS-TEMPERATURE Amount Off Rate Zero)
(assert-quantity BREAD-DARKNESS Amount Untoasted Rate Zero)
(assert-quantity OUTLET-CHARGE Amount On Rate Zero)
(assert-quantity GRAVITY Amount G Rate Zero)
(tick) ;t=1
(assert-quantity LEVER-POSITION Rate Negative)
(assert-quantity CARRIAGE-POSITION Rate Negative)
(tick) ;t=2
(assert-quantity LEVER-POSITION Amount Down Rate Zero)
(assert-quantity CARRIAGE-POSITION Amount Down Rate Zero)
(assert-quantity COILS-TEMPERATURE Rate Positive)
(tick) ;t=3
(assert-quantity BREAD-DARKNESS Rate Positive)
(tick) ;t=4
(assert-quantity COILS-TEMPERATURE Amount Hot Rate Zero)
(tick) ;t=5
(assert-quantity LEVER-POSITION Rate Positive)
(assert-quantity CARRIAGE-POSITION Rate Positive)
(assert-quantity COILS-TEMPERATURE Rate Negative)
(assert-quantity BREAD-DARKNESS Amount Golden Rate Zero)
(tick) ;t=6
(assert-quantity LEVER-POSITION Amount Up Rate Zero)
(assert-quantity CARRIAGE-POSITION Amount Up Rate Zero)
(tick) ;t=7
(assert-quantity COILS-TEMPERATURE Amount Off Rate Zero)
```

Figure 1: Observation of a Toaster

Figure 2 shows a causal model for the toaster which satisfies all the physical and causal constraints described above. Italicized text depicts hypothesized, unobservable quantities. Arcs depict hypothesized causal dependencies between quantities.

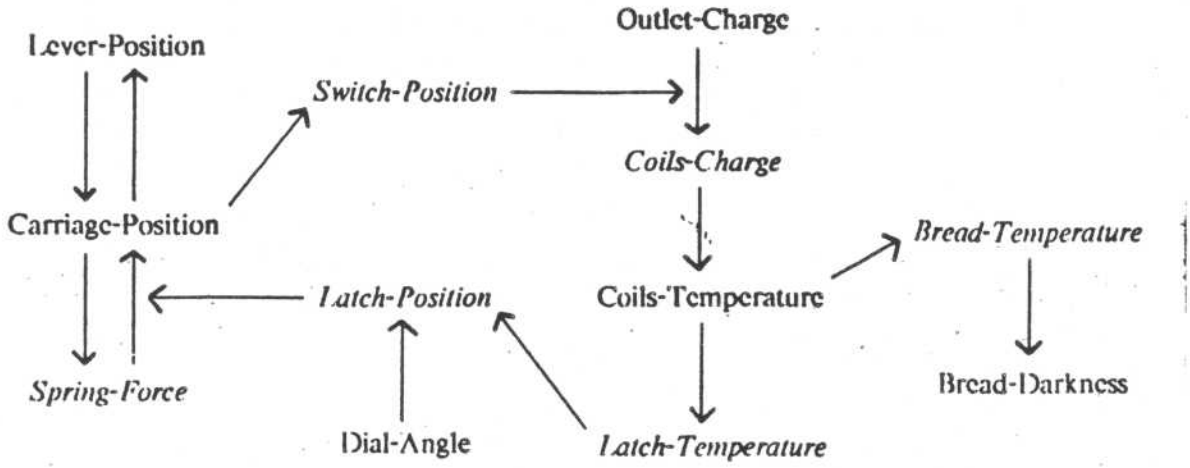

Figure 2: Hypothesized Configuration of Mechanisms for the Toaster

This example serves also to illustrate how reasoning about misbehavior can contribute to the causal modelling process. Consider an additional observation of the toaster in which the lever and carriage do not stay down. Further consider a failure schema for a broken latch which describes how the opposing surfaces of the latch never can overlap. The broken latch successfully accounts for the observed misbehavior and the hypothesis is retained. Similarly, a failure schema for a displaced latch whose surfaces cannot separate can explain a stuck lever and carriage which results in burnt toast.

An intriguing device is the pocket tire gauge. Most people quickly conjecture a mechanism within the cylinder to provide an opposing force to the air pressure so that a reliable reading is obtained. However, the slide which pops out of the cylinder remains right where it is when the gauge is removed from the tire stem and the force of the air pressure goes away. I leave the reader to ponder this puzzle while I try to get my causal modelling system to do the same.

Relation to Other Work

My work shares a methodological simlarity with Davis' work [Davis 84] on fault diagnosis. An effective troubleshooting system must entertain only reasonable hypotheses about faults. Davis is able to define reasonableness in terms of a set of general assumptions about models of physical systems which, when violated, lead directly to classes of faults. Similarly, I define reasonableness for design hypotheses in terms of a set of physical and causal constraints.

The recent work by de Kleer and Williams [de Kleer and Williams 87] on diagnosing multiple faults describes a minimalist representation for hypothesis spaces with a great degree of overlapping structure. This approach appears relevant to my problem.

I view my work as complementing Forbus' efforts [Forbus 84] at designing general representations for reasoning about physical systems. My representations are tailored towards the goal of exposing the physical and causal knowledge which constrains hypothesizing about hidden mechanisms.

The research effort probably most similar to mine is Shrager's work on *instructionless learning* [Shrager 85]. His research also investigates the creation of device models from observations of behavior. Shrager focuses on the process of model construction and refinement while my emphasis is on the knowledge which makes the problem tractable. In addition, his work has a cognitive science aspect.

Like Shrager's work, my work straddles the boundary between qualitative physics and machine learning. The recent work on explanation-based learning suggests a method for acquiring useful compositions of mechanisms to improve the performance of the causal modelling system. I have pursued this direction [Doyle 86].

Technology Transfer

Some of the results of the causal modelling project will be used in a joint project between the Ames Research Center and the Jet Propulsion Laboratory of NASA. The problem area is monitoring the behavior of complex physical systems through selected use of sensors. The proposed approach involves using causal models and qualitative simulation to generate expectations about the behavior of a physical system, then to plan the efficient use of sensors around selected checkpoints to verify proper functioning of the system. The same method is also appropriate for gathering evidence about fault hypotheses in a system which is misbehaving.

This project will utilize the representations for causality and the vocabulary of mechanisms being developed for the causal modelling research as the language for describing the physical systems to be monitored. The proposed problem domain is thermal management of the environments for the crew, cargo, and scientific experiments onboard our nation's proposed space station.

References

[Davis 84] Davis, Randall, "Diagnosis Based on Structure and Behavior," *Artificial Intelligence*, Vol. 24, pp. 347-410, 1984.

[de Kleer and Williams 87] de Kleer, Johan and Williams, Brian C., "Diagnosing Multiple Faults," *Artificial Intelligence, to appear*, 1987.

[Doyle 86] Doyle, Richard J., "Constructing and Refining Causal Explanations from an Inconsistent Domain Theory," *Proceedings of the National Conference on Artificial Intelligence*, Philadelphia, pp. 538-544, 1986.

[Forbus 84] Forbus, Kenneth D., "Qualitative Process Theory," *Artificial Intelligence*, Vol. 24, pp. 85-168, 1984.

[Shrager 85] Shrager, Jeffrey C., "Instructionless Learning: Discovery of the Mental Model of a Complex Device," Ph.D. diss., Carnegie-Mellon University, 1985.