A Set of Constraints for Reasoning in the Physical System Domain

Richard J. Doyle

Massachusetts Institute of Technology
545 Technology Square
Cambridge, Massachusetts 02139

Length: 3337 words plus 4 figures

Topic: Commonsense Reasoning (causal reasoning, qualitative reasoning)

Abstract: I describe a set of constraints which support reasoning about physical systems. These constraints are applied to the modelling or "black box" problem for devices: forming hypotheses about hidden mechanisms within devices from externally observable behavior. I relate in detail the performance of an implemented causal modelling system on an example involving the surprisingly puzzling pocket tire gauge. Results from several implemented examples indicate that this set of constraints supports capabilities for maintaining manageably sized hypothesis sets and for making fine distinctions among hypotheses.

A Scenario

The pocket tire gauge is a surprisingly puzzling device, despite its small range of behavior. If motion of the slide in a tire gauge is simply 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.

To get past this quandary, one has to note that there are couplings which allow motion in one direction but not in the opposite direction. One of these is a ratchet. However, once again observation does not provide confirmation. The slide may be pushed easily back into the cylinder when the gauge is off the tire. However, there is another kind of one-way coupling which is consistent with all of the observable behavior of the tire gauge. This is a coupling based simply on contact, not attachment, with which it is possible to push, but not to pull.

When the gauge is placed on a tire, released air enters the cylinder and pushes a piston inside the cylinder. This piston eventually touches and then pushes the slide. The piston is spring-loaded so that its motion is arrested when the restoring force of the spring balances the force due to the air pressure. The slide, no longer being pushed by the piston, also stops moving. When the gauge is removed from the tire, the force due to air pressure disappears and the now-unopposed spring pushes the piston back into the cylinder. However, the slide – unattached to the piston – stays right where it is. See Figure 1.

The design of the pocket tire gauge is elegant and proves obscure for most people. I have implemented a program called JACK which is able to achieve this modelling task.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the 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.

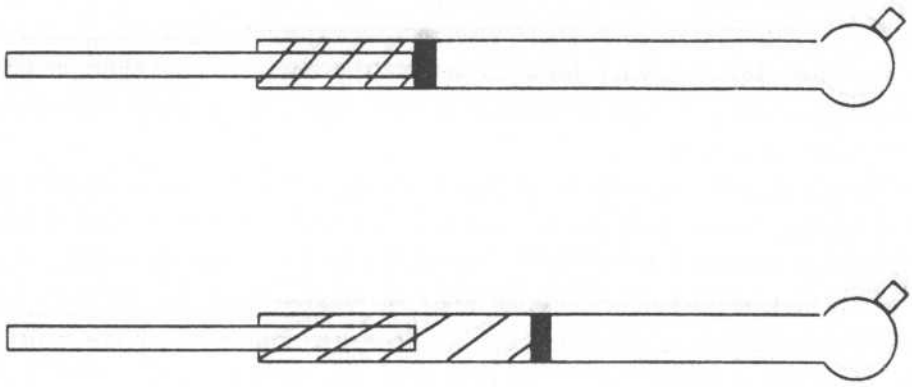

Figure 1. How a tire gauge works.

The Problem

Causal modelling is constructing causal explanations to account for the externally observable behavior of a device [Doyle 86, Doyle 88]. The task involves hypothesizing mechanism configurations inside the "black box" which are consistent with device behavior.

There are two inputs to the causal modelling system JACK: 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 accounting for the behavior of the device. See Figure 2.

The description of the behavior of a device consists of a timeline in which changes in the observable quantities of the device are recorded. For example, part of the description of the behavior of a tire gauge involves changes in the position of the slide. Initially, the slide is stationary. Some time later, it moves out of the cylinder, reaching a new stationary position. The slide does not move again.

Examples of mechanisms are mechanical couplings, thermal expansion, fluid flow, condensation, gravity, springs, valves, etc. These mechanisms serve as the primitive causal explanations from which the model of a device is constructed. They map causes to effects. For example, a mechanical coupling maps the motion of one physical object to the motion of another physical

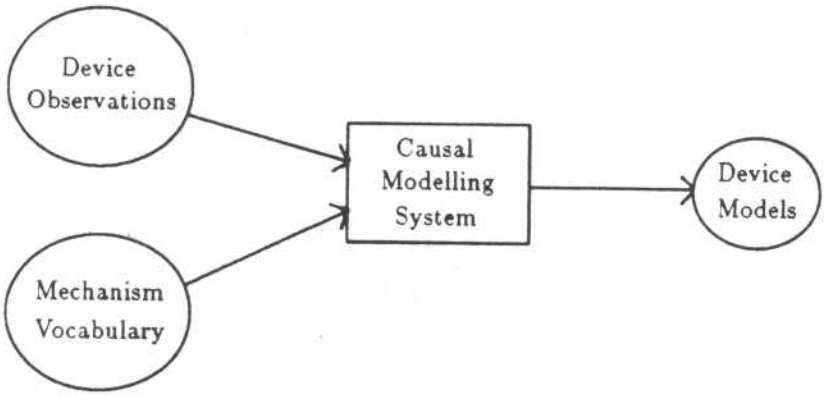

Figure 2. The causal modelling problem.

object. There are 50 mechanisms in the vocabulary of the causal modelling system.

The causal modelling task may be cast as a graph problem. The nodes of the graph correspond to device events - changes in the values of quantities. The arcs of the graph correspond to mechanisms which map events to other events. The set of observable events forms the periphery of graphs. The task is to construct a set of directed graphs which connect the observable event nodes. See Figure 3 and Figure 4. The direction of the arcs is from cause to effect. The mechanisms and intermediate events represent hypotheses about what hidden mechanisms may exist and what unobservable events may take place inside the black box.

The causal modelling problem is hard because in the worst case there is a doubly exponential number of possible hypotheses about what mechanisms may be inside a device.

For any pair of events, the number of possible mechanism paths between them is $\mathcal{O}(m^l)$ where m is the number of possible mechanisms and l is the length of the path. Unfortunately, linear mechanism chains are not the only form of hypothesis which must be considered. An effect may be the result of an interaction between multiple causes. The *equilibrium* state in the tire gauge is an example. The number of hypotheses for interacting, or joined mechanism paths is the product of the number of hypotheses for each of the separate linear

mechanism paths. For p interactions, the number of mechanism hypotheses becomes $\mathcal{O}(m^{lp})$.

The Domain

The domain of investigation for this work is mechanical, electrical, thermal, and pneumatic physical systems. This class does not include electronic devices: digital, analog, or VLSI technology. The order of complexity which has been tackled is roughly that of the common household gadget.

The device examples which have been implemented include a toaster, a tire gauge, an old-style bicycle drive with coaster brake, a refrigerator, and a home heating system. The program JACK models simplified versions of the more complex among these physical systems.

The Approach

My approach to making the causal modelling problem tractable is a three-pronged approach. One of the prongs involves applying a set of constraints which embody physical and causal principles to prune hypotheses. Another prong involves enumerating different forms for hypotheses, placing an ordering on these forms, and using this ordering to carefully control the generation of hypotheses. A third prong involves a straightforward use of abstraction spaces. The pruning power resulting from the combined application of these three thrusts has proven to be impressive. Because of space limitations, only the physical and causal constraints are treated in this paper.

A Set of Constraints

The constraints support reasoning about how mechanisms map device inputs to device outputs. Each constraint concerns a different observable aspect of the behavior and structure of physical systems. All hypotheses about hidden mechanism configurations within devices must account for any observed changes or lack of changes between cause events and effect events for all of these aspects of behavior and structure.

The *type* constraint concerns the types of quantities in physical systems. Hypotheses must account for observed type conservations or transformations

between causes and effects. For example, 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 physical systems. Hypotheses must account for observed time lags between causes and effects. For example, electricity or a rigid coupling, whose propagation times are essentially instantaneous, are consistent hypotheses for a cause and effect which are perceptually simultaneous. Conversely, these same mechanisms cannot be offered as an explanation for events which are separated in time.

The *sign* constraint concerns the signs of the values of quantities in physical systems. Hypotheses must account for any change or lack of change of sign between causes and effects. For example, an increase in temperature can account for an increase in pressure but cannot explain a decrease in pressure. Flow in a closed system implies a decrease in amount at the cause and an increase at the effect, or vice versa. In an open system, both amounts may increase or both amounts may decrease.

The *direction* constraint concerns the orientations of quantities in physical systems. Hypotheses must account for any deflections between causes and effects. The direction constraint is an elaboration of the sign constraint for vector, as opposed to scalar, quantities. 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. A rigid coupling, on the other hand, which preserves orientation, is not.

The *magnitude* constraint concerns the magnitudes of the values of quantities in physical systems. Hypotheses must account for any decreases, increases, or lack of change in magnitude between causes and effects. For example, a rigid coupling, which transfers motion with no loss, can be a causal explanation only for motions of the same magnitude. Gravity can account for finite motions only within a certain range of magnitude, even given the effects of acceleration.

The *alignment* constraint concerns the relative values of quantities in physical systems. Hypotheses must incorporate any inequality relations imposed by mechanisms between causes and effects. For example, the direction of heat flow always is from the warmer to the cooler site. Or, stated differently, the temperature value at the cause must be greater than the temperature value at the effect. This constraint also distinguishes couplings which support pulling but not pushing, or vice versa. For example, 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 physical systems. Hypotheses must incorporate any restrictions concerning absolute directions of change imposed by mechanisms between causes and effects. For example, a ratchet allows motion in one direction but not in the opposite direction. A coupling based on contact, on the other hand, may engage in either direction. Condensation results from a pressure increase and evaporation results from a pressure decrease.

The *displacement* constraint concerns the locations of objects in physical systems. Hypotheses must account for any physical separation between causes and effects. For example, 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. However, thermal expansion preceded by a heat flow, or thermal expansion followed by a mechanical coupling can explain the observation because in both cases, the additional mechanism is sufficient to account for the change in location.

The *medium* constraint concerns the structural connections between objects in physical systems. Only those hypotheses for which the appropriate structural connections between causes and effects can be established or conjectured may be admitted. For example, gas flow is an admissible hypothesis when two physical objects are joined, but is untenable when they are separated. A valve must span a conduit in order to explain a change in flow.

The Tire Gauge Example

One of the tasks set for the program JACK in the tire gauge example is to explain why the slide stops moving before reaching its limit position. The hypothesis which corresponds to the way a real tire gauge works appears in Figure 3. Here the causal modelling system conjectures that an equilibrium has been achieved. The two opposing contributions which make up the equilibrium are a pneumatically-induced motion of a hidden physical object due to the flow of gas from the tire, and a spring-induced motion of the same physical object due to displacement of a spring by the moving object, which results in a restoring force in the direction opposite to the displacement.

A note concerning the figures: Observable events are denoted by solid circles and are described in terms of a physical object, a quantity type, a value, and a moment. Conjectured events are denoted by open circles. Mechanisms are denoted by solid arcs. Dotted arcs denote temporal integration episodes during which the value of a quantity changes.

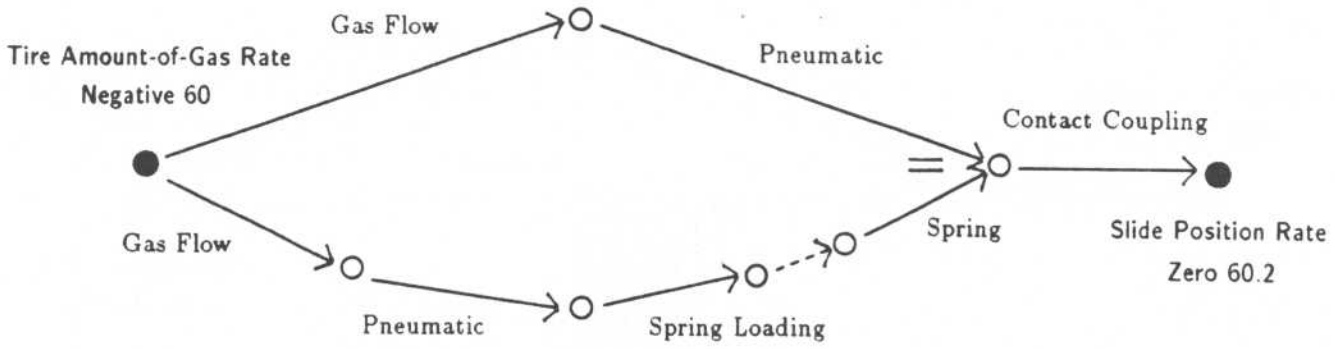

Figure 3. Spring hypothesis.

The triggering heuristics for suspecting equilibrium and disablement situations are the same: an unexpected zero value occurs after an expected non-zero effect. Not surprisingly, the program JACK is able to generate hypotheses involving disablement to explain the halting of the motion of the slide. One of these hypotheses appears in Figure 4. This proposed causal model for the tire gauge also involves pneumatically-induced 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 stops, and the motion of the slide - transmitted along a mechanical coupling from the hidden object - also stops. Thus an impulse of displaced gas is conjectured to be responsible for the start-and-stop motion of the slide.

An alternate disablement hypothesis generated by the causal modelling system proposes that the pneumatic motion of the hidden object, rather than closing a valve which disables the flow of gas, instead engages a latch which directly arrests the motion of the slide.

Another opportunity to reason about the spring and impulse hypotheses is afforded by the part of the tire gauge observation which describes how the slide, which had been stationary, continues to be motionless when the cylinder of the tire gauge is removed from the tire. This part of the observation, while not distinguishing the two hypotheses, does shed some light on the nature of the mechanical coupling between the hidden object and the slide conjectured in both hypotheses.

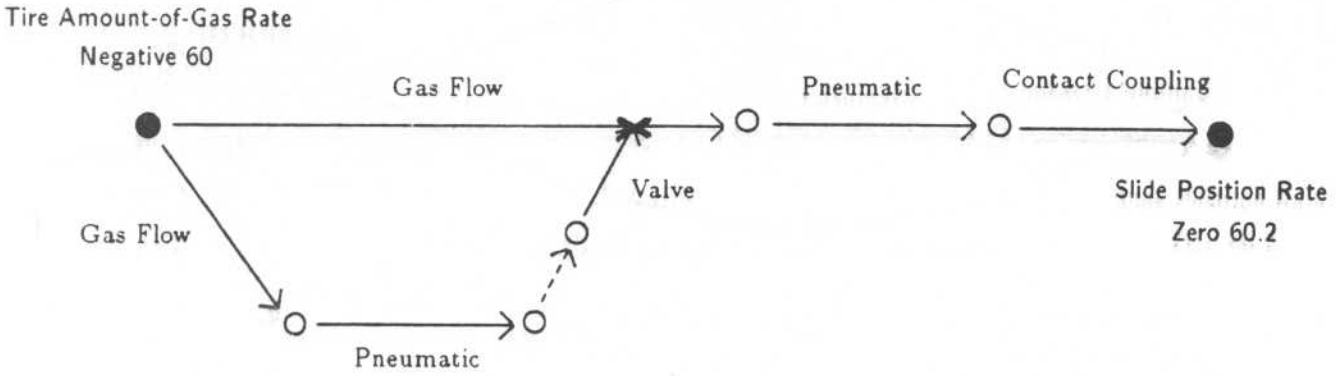

Figure 4. Impulse hypothesis.

The relation {*Tire Joined-To Cylinder*} now is false and violates the medium constraint for the *Gas-Flow* mechanism in both hypotheses. For the impulse hypothesis, this results in both the primary path and the disabling path in a disablement interaction becoming inactive. A prediction of no expected effect is consistent with the observation of the slide remaining motionless.

The reasoning for the spring hypothesis is considerably more subtle. Both halves of the proposed equilibrium become inactive because the now-unsupported *Gas-Flow* mechanism appears as the first mechanism along both interacting paths. However, and this is a key point, the two mechanism paths do not become inactive at the same time. The delay along the mechanism path which contains the spring is longer. Just as time is required to displace the spring and achieve the equilibrium state, so time is required to unload the spring and remove this influence on the position of the hidden object. There is an interval during which the pneumatic half of the equilibrium interaction has become inactive while the spring half is still active. The program JACK is able to infer this broken equilibrium from the unequal delays along the two mechanism paths and predicts that the hidden object moves in the direction opposite to its original motion.

The task now is to explain how the slide need not move despite the conjectured motion of the hidden object inside the tire gauge. Three types of mechanical couplings between the hidden object and the slide are proposed

by the causal modelling system in both the spring and the impulse hypotheses: the *Rigid-Coupling*, *Contact-Coupling*, and *Ratchet* mechanisms. The *Rigid-Coupling* is predicted to be active and is inconsistent with the motionless slide. The slide should move into the cylinder along with the hidden object. The *Contact-Coupling* mechanism is predicted to be inactive because the alignment constraint is violated: the position of the hidden object is greater than, not less than, the position of the slide along the direction of motion. In other words, the hidden object cannot pull the slide. This mechanism can explain the stationary slide. The *Ratchet* mechanism also is predicted to be inactive because the bias constraint is violated: the motion is not in the only direction allowed. This mechanism also is compatible with the slide remaining at rest.

The *Ratchet* mechanism ultimately is eliminated when the slide is pushed back manually into the cylinder. This observation is inconsistent with the prediction that the slide will not move in this direction.

Empirical Results

Research efforts in artificial intelligence must be evaluated on two criteria: the generality of the principles articulated in the work, and the computational utility of those principles.

I have outlined the character of the principles embedded in the causal modelling system JACK and described the diversity of the reasoning supported by those principles in the context of the tire gauge example. In this section, I offer empirical results concerning the pruning power inherent in those principles.

Table 1 shows the number of hypotheses admitted for each of the implemented device examples. l_{max} is the length of the longest mechanism path in any hypothesis for the given example. p_{max} is the greatest number of interacting mechanism paths in any hypothesis for the given example. Hypothesis refinement over multiple instances of behavior was disabled in these runs; the concern here is to determine the size of the initial set of hypotheses produced.

Device	l_{max}	p_{max}	*Hypotheses*
Toaster	2	2	28
Tire Gauge	5	2	103
Bicycle Drive	2	2	4
Refrigerator	4	2	263
Home Heating	3	3	517

Table 1. Number of hypotheses admitted.

The overall pruning ratios achieved are impressive. In the case of the tire gauge, the worst case number of hypotheses given a vocabulary of 50 mechanisms is on the order of $50^{(5\times2)} \approx 10^{17}$.

The results in Table 1 also reflect the pruning contributions of an ordering on hypotheses and of abstraction spaces. These secondary sources of pruning power are not discussed in this paper; they contribute approximately three or four additional orders of magnitude to the pruning ratios.

Relation to Other Work

Causal and qualitative simulation plays a role in modelling. Device hypotheses are simulated by propagating values for each of the constraints. Predictions are compared to observed events and form the basis for admitting or pruning hypotheses.

Several approaches to causal and qualitative simulation have appeared in the literature. Seminal works among these include Forbus' Qualitative Process Theory [Forbus 84], de Kleer and Brown's qualitative physics based on confluences [de Kleer and Brown 84], and Kuipers' method for inferring behavior from causal structure [Kuipers 84].

The set of constraints described in this paper support a complementary approach to causal and qualitative simulation. Representing the behavior and structure of physical systems in terms of this set of constraints supports reasoning about which changes occur, what new values are reached, what are the times and locations of events, which mechanisms are active and which are inactive, and which interactions occur.

Shrager, in his research on instructionless learning [Shrager 87], also investigates the modelling problem. Shrager focuses on a cognitive model of

device hypothesis construction and refinement in humans while my emphasis is on the sources of constraint which make the problem tractable.

Causal modelling can be cast as an instance of Waltz network labelling [Waltz 75]: The networks are the causal graphs which represent hypotheses about hidden configurations of mechanisms within a device. The arcs are labelled with mechanisms and the nodes are labelled with values for the constraints which describe events.

A singular difference separates the causal modelling problem from other instances of Waltz labelling - the network is not known. Only the peripheral nodes and their labellings are known. These are the externally observable events. During the causal modelling process, networks are constructed by conjecturing mechanism arcs and additional event nodes.

The performance of the program JACK offers an extraordinarily convincing demonstration of the potential power of the Waltz network labelling technique. In the right domain and with the right constraints, the network need not even be known. Network topologies can be generated in concert with the actual labelling process.

Conclusions

Exposing sources of constraint is part of Marr's well-known methodology for conducting research in artificial intelligence [Marr 82]. Constraint sources which are amenable to tidy representation can enable otherwise prohibitively large hypothesis spaces to be searched effectively. The best constraints focus away from overwhelming detail while retaining discriminatory power.

I have identified a set of constraints for reasoning about various aspects of the behavior and structure of physical systems. I have applied these constraints to the difficult modelling or "black box" problem for devices. Results from several implemented examples indicate that this set of constraints supports capabilities for maintaining manageably sized hypothesis sets and for making fine distinctions among hypotheses.

References

[de Kleer and Brown 84] de Kleer, Johan, and John S. Brown, "A Qualitative Physics Based on Confluences," *Artificial Intelligence* **24**, 1984.

[Doyle 86] Doyle, Richard J., "Constructing and Refining Causal Explanations from an Inconsistent Domain Theory," *National Conference on Artificial Intelligence*, Philadelphia, 1986.

[Doyle 88] Doyle, Richard J., "Opening up the Black Box: Hypothesizing Mechanisms Within Devices," Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, Mass., *in preparation.*

[Forbus 84] Forbus, Kenneth D., "Qualitative Process Theory," *Artificial Intelligence* **24**, 1984.

[Kuipers 84] Kuipers, Benjamin P., "Commonsense Reasoning about Causality: Deriving Behavior from Structure," *Artificial Intelligence* **24**, 1984.

[Marr 82] Marr, David, *Vision*, W. H. Freeman and Company, New York, 1982.

[Shrager 87] Shrager, Jeff, "Theory Change via View Application in Instructionless Learning," *Machine Learning* **2**, no. 3, 1987.

[Waltz 75] Waltz, David, "Understanding Line Drawings of Scenes with Shadows," in *The Psychology of Computer Vision*, P. Winston (ed.), McGraw-Hill, New York, 1975.

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