Physical Idealization as Plausible Inference

Ernest Davis*
Courant Institute
New York, New York

May 25, 1994

Abstract

The analysis of physical systems almost always relies on idealized models of the objects involved. Any idealization, however, will be incorrect or insufficiently accurate some of the time. It seems reasonable, therefore, to view a physical idealization as a defeasible inference which can be withdrawn in the presence of contrary evidence. This talk discusses the consequences of such a view.

We focus on examples where a system may or may not go into a state where idealizations are violated, such as dropping a ball near an open switch connected across a battery. We show that:

1. Non-monotonic logics will try to enforce the idealization by supposing that the ball will miss the switch. This anomaly does not seem to be solvable by the kinds of techniques that have been applied to the Yale Shooting Problem, which it superficially resembles. We show that this problem is analogous to anomalies in non-monotonic logic that are time-independent.

2. A probabilistic analysis is possible, but it relies on independence assumptions that are hard to justify in general.

3. For completely specified systems, the rule "If the idealization gives solvable equations, then assumes that it holds" is, in fact, a monotonic system of inferences. It should therefore be possible to characterize this in a purely deductive theory. We show that this is, indeed, possible for simple cases, but can get messy in complex systems.

4. Programs that make physical predictions can avoid these problems by simply avoiding reasoning from the future to the past. Though most current programs observe this restriction, it seems likely that more powerful and general systems will have to violate it, and thus deal with this issue.

5. Finally, we look at dynamic systems where the idealization can be observed at any single instant, but it is inconsistent over extended intervals.

Reasoning about physical systems almost always relies on idealized models of the objects involved. Strings and pulleys are massless; springs obey Hooke's law; solid objects are perfectly rigid; and so on [Addanki et al., 89]. Generally, analyses based on such idealizations are correct qualitatively and reasonably accurate quantitatively. (What constitutes reasonable accuracy depends both on the requirements of the task for which the analysis is being performed and on the availability of alternative models that are both more accurate and tractable. In practice, modellers often use idealizations that are known to be inaccurate, either because the next more accurate model is horribly complex, or because no more accurate model is known.)

Sometimes, however, simple idealizations give answers that are not adequate. In this case, the modeller must either use more accurate, presumably more complex, models, or abandon modelling

*This research was supported by NSF grant #IRI-9001447. Thanks to R. Bhaskar, Hector Geffner, Benjamin Grosof, Leo Joskowicz, Leora Morgenstern, and Wlodek Zadrozny for helpful comments and criticisms.

altogether and resort to physical experimentation. A modeller can detect the breakdown of an idealization in any of the following ways:

1. It may be known that the circumstances lie outside the range in which the idealization is accurate. For example, a watch spring asked to support a ton weight will not obey Hooke's law.

2. It may be known that more accuracy is required than the idealization provides. For example, in most cases, it can be assumed that electronic signals propagate instantaneously. In the design of a Cray, however, where the diameter of the device is substantial as compared to the cycle time of the machine times the speed of light, such an approximation cannot be made.

3. Physical measurements may show that the actual behavior differs significantly from the predicted one.

4. The ideal model may predict behavior that is widely at variance with a commonsense knowledge of physics. For example, an idealized model of a pool table, with or without contact friction, would predict that a ball can never stop rolling in the middle of a table. This contradicts the commonsense knowledge that rolling objects slow down and stop in finite time.

5. The ideal model may predict behavior that is inconsistent with known global constraints. Examples: Designs for perpetual motion devices.

6. The ideal model may be internally inconsistent. An example is shown in figure 1: A battery is short-circuited by a wire. If the battery is modelled as a voltage source, which always maintains a constant voltage difference between its two ends, and the wire is modelled as a short circuit, which always maintains equal voltage at its two ends, then the mathematical model involves the following inconsistent pair of equations:

   $V_A - V_B = V_0 > 0$. (Battery)
   $V_A = V_B$ (Wire)

From a logical point of view, (4), (5), and (6) can all be characterized as the model being inconsistent with the modeller's physical knowledge, where this is taken to include both formal physics, as in (5), and commonsense physical knowledge, as in (4). Problems of type (3) can likewise be subsumed in this category, if the contradictory measurements are taken to be part of the modeller's physical knowledge. In this paper, we will confine attention to problems in these categories. Whether there is a natural class of problems where problems of categories (1) and (2) can be ignored — that is, where it can be assumed that the idealizations are either reasonable or impossible — is a difficult issue that we will here ignore.[1]

Consider, again, the example in figure 1, of a battery short-circuited by a wire. If we took the idealizations of the components to be universally valid rules, "A battery always maintains a voltage difference between its poles," and "A wire always maintains equal voltage at its poles," then the only conclusion would be that this circuit cannot be built.[2] There are cases where such a

[1] A.E. Housman has some scathing remarks on this assumption, in the context of editing literary texts: "To believe that wherever a best manuscript gives possible readings it gives true readings, and that only where it gives impossible readings does it give false readings, is to believe that an incompetent editor is the darling of Providence, which has given its angels charge over him lest at any time his sloth and folly should produce their natural results and incur their appropriate penalty. Chance and the common course of nature will not bring it to pass that the readings of a MS. are right wherever they are possible and impossible wherever they are wrong; that requires divine intervention; and when one considers the history of man and the spectacle of the universe I hope one may say without impiety that divine intervention might have been better employed elsewhere." (Preface to Manilius I)

[2] The inconsistency here cannot be associated with the advice that the circuit *should* not be built. There is no connection between violating mathematical idealization and unwise design.

conclusion is correct; for example, a description of a mechanical system that specifies that a two-foot rod is connected at both ends to a one-foot rod. With the battery example, however, we wish to conclude, instead, that the physical idealizations do not apply. Either the battery does not succeed in maintaining the voltage difference; or the wire has a certain resistance; or the system radically changes before reaching equilibrium (for instance, the wire melts).

Let us make the simplifying assumption, for the moment, that the appropriate fix here is to associate a small resistance with the wire. Thus, in most cases, we wish to approximate a wire as a short circuit; here, we wish to approximate it as a resistor of very low resistance. One way to accomplish this would be to change the basic model of a wire to be a small resistor; to apply this uniformly to all circuits; and then to use mathematical techniques to justify ignoring this resistance in almost all cases. (One might, for example, treat the resistance of a wire as an infinitesimal quantity, and use techniques like those of [Raiman, 86] to show that its effect is negligible.) But the computational costs of this approach are non-trivial, and become worse as more tiny but non-zero characteristics are found to be occasionally important. For example, a physical wire has non-zero capacitance and inductance; and it is possible to construct systems where these have non-negligible effects. If we include these as very small, but non-zero, quantities in our analysis of all models that contain wires, the effect is to add a second-order differential equation for each wire in the system, and then to remove it through a perturbation analysis. Pursuing this further, one would end up setting up Maxwell's equations, or Schrodinger's equations, or quantum electrodynamics for all space, and then throwing out all the lower-order terms. This is not feasible. Rather, reasoners will proceed by invoking the idealized assumption that a wire supports no voltage difference as a plausible inference at the time when a mathematical model is constructed from the given physical description, and replacing it by a more accurate assumption if necessary.

Thus, given a physical description that specifies that there is a wire between nodes A and B, we will prefer to model this by the equation $V_A = V_B$ if this is possible; if it is not, we will use a more detailed model.

The remainder of the paper studies theories that incorporate such inferences. Section 1 exhibits a simple example that is difficult to handle correctly if physical idealization is treated as a default rule in a non-monotonic logic. Section 2 show that probability theory can be made to give the right answer to the problem, but doing so involves some anomalies. Section 3 discusses the possibility of using deductive rather than defeasible inference. This can be made to work for simple microworlds, but it relies on an implicit notion of causal directionality that is, itself, hard to formalize. Section 4 discusses the relevance of these difficulties to physical prediction programs. Section 5 discusses a variety of examples where the failure of the idealization can be detected only over an extended interval, not at a single instant; such examples are even more difficult to analyze.

1 Non-monotonic logic

At first glance, physical idealizations would seem a natural domain for non-monotonic logic. We prefer to use the idealization if it is consistent; we drop it only if it is inconsistent. For example, we could have a rule stating, "If A and B are connected by a wire, then assume, barring contradictory information, that there is no voltage difference between the two." Thus, in most circuits, this conclusion could be drawn for all the wires; in the circuit of figure 1, the conclusion would be blocked, since it leads to an contradiction.

Unfortunately, there are physical systems where reasonable conclusions cannot be drawn if idealizations are attached to non-monotonic rules. Consider the circuit in figure 2; a battery is attached to a switch. The natural idealization of a switch is that it acts as an open circuit (zero current flow) if it is open and as a short circuit (zero voltage difference) if it is closed. Suppose that we do

not know whether the switch is open or closed. If we consider the idealizations of the battery and the switch to be defaults in a non-monotonic logic, then the logic will force the conclusion that the switch is open. Since it is possible that both defaults are satisfied, we must assume that they are both satisfied; and if they are both satisfied, the switch must be open.

This conclusion seems odd. It becomes even odder if the uncertainty about the state of the switch rests on uncertainty about the outcome of previous events. Suppose that the switch is originally open, and we drop a ball somewhere nearby (Figure 3). If the ball falls on the switch, it will close; otherwise, it remains open. Our information about the ball is not precise enough to determine which will happen. Now, if the switch closes, one of our idealization defaults will have to be violated; if it remains open, they can all be maintained. Since it is consistent that they can all be maintained, the non-monotonic logic requires the conclusion that this is what happens; therefore, the ball will miss the switch.

This anomaly looks something like the Yale Shooting Problem [Hanks and McDermott, 87]. In both cases, we are reasoning backwards from a non-monotonic inference applied in the future to a conclusion about the past. However, our anomaly here has nothing to do with the frame problem, and, in fact, does not depend in its logical form on temporal issues at all. Note, also, that in the Yale Shooting Problem, the result generated by the logic is weaker than desired; in our problem, it is stronger.

1.1 Penguins and pelicans: a non-monotonic anomaly

To bring this analysis into a more familiar framework, we show an analogue within the paradigm non-monotonic example of birds that fly. Suppose we have the following theory:

General Rules:
1. All penguins are birds.
2. All pelicans are birds.
3. No penguins can fly.

Default Rule:
4. Birds can typically fly.

Problem Specification:
5. Tweety is either a pelican or a penguin.
6. Polly is a bird.

Intuitively, it should not be possible to decide whether Tweety is a penguin and cannot fly or is a pelican and can fly. However, in a straightforward application of non-monotonic logic, it will be possible to conclude that Tweety is a pelican. Since Tweety is certainly a bird, and it is consistent that he can fly, the default rule will support the conclusion that he can fly. From (3) it follows that he is not a penguin.

A solution to this anomaly would consist in a representation for the problem and a formulation of non-monotonic logic that allows the inference of "Polly can fly," but not of "Tweety can fly." The problem is that within the categories available in most well-known non-monotonic logics, the possibility that Polly is a penguin has exactly the same status as the possibility that Tweety is a penguin; they are both consistent but not necessary and they both violate one default rule.

Intuitively, the difference between (5) and (6) is that (5) suggests that Tweety being a pelican and Tweety being a penguin are both reasonable possibilities, while (6) makes no such suggestion. However, it is certainly not in general reasonable to infer "p is a reasonable possibility" from "p or q". Aside from anything else, one always knows "p or not p." It is not possible to rule out this case as trivial; the disjunction in our original example is, in fact, the tautology, "The ball either hits

the switch or does not hit the switch." The difference between this tautology and all other similar tautologies, such as "Either Smokey the Bear will flip the switch or he will not," is that we have some reason to think it fairly likely that the ball will hit the switch. But this concept of "reasonable likelihood" is difficult to capture in a non-monotonic logic.

In motivated action theory [Stein and Morgenstern, 89], a distinction is drawn between disjunctions of event occurrences that are specifically stated as part of a problem description, and those that are inferred from the general rule of excluded middle. If a problem specifically states that "Either the ball hits the switch or it misses the switch," the theory considers both events to be "motivated" and it therefore draws the correct conclusion that either may be possible. This interpretation of disjunctions can be used only with disjunctions of event occurrences, not with general propositions. Thus, MAT solves the ball and switch problem but not the penguin/pelican example. Motivated action theory, however, has been formulated for a discrete description of events; it is not clear whether it can be carried over into a more general physical domain where change is largely characterized by differential equations.

There are other ways in which non-monotonic logics can express the concept of being possible, but not necessary. In some logics, a possible fact can be viewed as one that is true in some extension, or which satisfies some default. Thus, if we represent "Tweety is a penguin or a pelican" by the pair of default rules "By default, Tweety is a penguin" and "By default, Tweety is a pelican" then we get more or less what we want. In default logic [Reiter, 80] there is one extension in which Tweety is a penguin and cannot fly and another extension in which Tweety is a pelican and can fly. In circumscription [McCarthy, 80] it is impossible to conclude that Tweety is a penguin. However, such a non-standard use of default rules seems to me perverse and dangerous, though some of the mavens in the area whom I have consulted seem to be willing to consider it, or analogous techniques in other formalisms, as a possible representation.

Another possibility would be to extend non-monotonic logic by an operator "There is some reason to believe ϕ" and then interpret a default rule, "If p then assume q," as "If p holds and there is no reason to believe $\neg q$, then infer q." (Hector Geffner informs me there has been some recent work in this direction, but I have not seen details.)

Note that if [3] and [5], above are reworded as "If a bird can fly, then it is not a penguin," and "If Tweety is not a penguin, then it is a pelican," the conclusion "Tweety is a pelican," seems much more palatable. The English phrase "not p implies q" does not suggest that both p and q are possibilities with anything like the same force as the logically equivalent "p or q".

2 Probability theory

The theory of probability, by contrast, explicitly deals with varying degrees of certainty, so we may expect that it can handle this problem. We will interpret the English sentence, "Tweety is either a penguin or a pelican," as meaning that the prior probability of each is .5. (All the numbers used in this example are fudge factors; the same kinds of results can be gotten with other numbers, or with constraints like "The probability that Tweety is a penguin is at least 0.1.") The formalization, however, is a little tricky. A first attempt might go along these lines:

1. $\forall_X$ penguin$(X) \Rightarrow$ bird(X)
2. $\forall_X$ pelican$(X) \Rightarrow$ bird(X)
3. $\forall_X$ penguin$(X) \Rightarrow \neg$can_fly(X)
4. $\forall_X$ P(can_fly(X) | bird$(X)) = 0.9$

5. $P(\text{penguin}(\text{tweety})) = P(\text{pelican}(\text{tweety})) = 0.5$

6. $\text{bird}(\text{polly})$

7. $\forall_X \neg[\text{penguin}(X) \wedge \text{pelican}(X)]$

However, this is inconsistent. Proof: From (1), (2), (5), and (7) it follows that $P(\text{bird}(\text{tweety})) = 1$. We then get

8. $P(\text{can_fly}(\text{tweety})) \geq P(\text{can_fly}(\text{tweety}) \wedge \text{bird}(\text{tweety})) =$
   $P(\text{can_fly}(\text{tweety}) \mid \text{bird}(\text{tweety})) \cdot P(\text{bird}(\text{tweety})) =$ (from (4))
   $0.9 \cdot 1.0 = 0.9.$

However, from (5) and (7) it also follows that there are two disjoint possibilities: either Tweety is a penguin or he is a pelican. Therefore

9. $P(\text{can_fly}(\text{tweety})) =$
   $P(\text{can_fly}(\text{tweety}) \wedge \text{pelican}(\text{tweety})) + P(\text{can_fly}(\text{tweety}) \wedge \text{penguin}(\text{tweety}))$

The event $\text{can_fly}(\text{tweety}) \wedge \text{penguin}(\text{tweety})$ is inconsistent with rule (3), and therefore has zero probability. We have, then,

10. $P(\text{can_fly}(\text{tweety})) = P(\text{can_fly}(\text{tweety}) \wedge \text{pelican}(\text{tweety})) =$
    $P(\text{can_fly}(\text{tweety}) \mid \text{pelican}(\text{tweety})) \cdot P(\text{pelican}(\text{tweety}))$

By (5), $P(\text{pelican}(\text{tweety})) = 0.5$, so $P(\text{can_fly}(\text{tweety})) < 0.5$.

The problem is that the probabilities in (4) and (5) are based on different states of knowledge. The probability in (5) that Tweety has a 50-50 chance of being a penguin or a pelican must be based on some observation α; for example, the observation that Tweety lives in an aviary where half the birds are penguins and the other half are pelicans. The probability in (4) applied to Tweety can only be understood as holding before the observation α is made. Thus, we must change (5) to read

5. $P(\text{penguin}(\text{tweety}) \mid \alpha) = P(\text{pelican}(\text{tweety}) \mid \alpha) = 0.5$

For symmetry, we will also assume that our knowledge that Polly is a bird rests on an observation β.

6. $P(\text{bird}(\text{polly}) \mid \beta)$

We are now interested in the probabilities that Tweety and Polly can fly, given the two observations α and β.

We begin with $P(\text{can_fly}(\text{tweety}) \mid \alpha \wedge \beta)$ First, we eliminate β from the conditions by assuming that $\text{can_fly}(\text{tweety})$ is independent of β given α. Thus we have

11. $P(\text{can_fly}(\text{tweety}) \mid \alpha \wedge \beta) = P(\text{can_fly}(\text{tweety}) \mid \alpha)$

From [5] and [7], there are two disjoint possibilities: either Tweety is a penguin or he is a pelican. Therefore

12. $P(\text{can_fly}(\text{tweety}) \mid \alpha) =$
    $P(\text{can_fly}(\text{tweety}) \wedge \text{pelican}(\text{tweety}) \mid \alpha) + P(\text{can_fly}(\text{tweety}) \wedge \text{penguin}(\text{tweety}) \mid \alpha)$

The event can_fly(tweety) $\wedge$ penguin(tweety) is inconsistent with rule (3), and therefore has zero probability. We have, then,

13. P(can_fly(tweety) | α) = P(can_fly(tweety) $\wedge$ pelican(tweety) | α) =
P(can_fly(tweety) | pelican(tweety) $\wedge$ α) $\cdot$ P(pelican(tweety) | α)

The second factor in the last line is 0.5, by (5). We can evaluate the first factor as follows: First, we assume that can_fly(tweety) is independent of α given pelican(tweety). From this we conclude that

14. P(can_fly(tweety) | pelican(tweety) $\wedge$ α) = P(can_fly(tweety) | pelican(tweety))

Next we observe that bird(tweety) is a consequence of pelican(tweety) and (2). Hence

15. P(can_fly(tweety) | pelican(tweety)) = P(can_fly(tweety) | bird(tweety) $\wedge$ pelican(tweety))

Finally, we assume that, given the event bird(tweety), the event can_fly(tweety) is independent of the event pelican(tweety). Then we have

16. P(can_fly(tweety) | bird(tweety) $\wedge$ pelican(tweety)) =
P(can_fly(tweety) | bird(tweety)) = 0.9.

Combining (11-16), we conclude

17. P(can_fly(tweety) | $\alpha \wedge \beta$) = $0.9 \cdot 0.5 = 0.45$.

Note that the calculation above makes three independence conditions:

IND.1 The event can_fly(tweety) is independent of the observation β. This is reasonable, since β is an observation of Polly that should not affect one's beliefs about Tweety.

IND.2 The event can_fly(tweety) is independent of α given pelican(tweety). That is, all that observation α tells us relevant to whether Tweety can fly is that he is either a penguin or a pelican, with a fifty-fifty chance of each.

IND.3 The event can_fly(tweety) is independent of pelican(tweety) given bird(tweety). This is the least secure of our independence assumptions. Note, first, that we cannot substitute "penguin" for "pelican" above; it is not consistent to suppose that can_fly(tweety) is independent of penguin(tweety) given bird(tweety). (Finding out that Tweety is a penguin changes the probability that Tweety can fly from 0.9 to 0.0.) It is hard to find a justification for this assumption other than the weak rule, "For any events α, β and ϕ, it may be assumed, in the absence of other indications, that α is independent of ϕ given β."

Another difficulty is that "can_fly(X)" is probably not actually independent of "pelican(X)" given "bird(X)." Other things being equal, one would expect that the incidence of flying ability is higher among pelicans than among birds as a whole, if only because all the non-flying penguins have been removed from the sample space.[3] That, of course, strengthens our argument that Tweety probably can fly, but it does require a slightly changed analysis. We can formalize this as follows:

[3] A richer theory might specify that pelicans form a species; that, in most cases, either virtually all the members of a species can fly or none can; and that 90% of bird species can fly. This would change the form of the analysis, but the conclusion would end up the same. In particular, the probability that pelicans can fly would still be slightly raised by the knowledge that penguins cannot fly.

Given that P(can_fly(X) | bird(X)) = 0.9 > 0 = P(can_fly(X) | penguin(X) $\wedge$ bird(X)), we can infer that P(can_fly(X) | bird(X) $\wedge$ $\neg$penguin(X)) > P(can_fly(X) | bird(X)) = 0.9. Now, if we assume that can_fly(tweety) is independent of pelican(tweety) given [bird(tweety) $\wedge$ $\neg$penguin(tweety)], we can calculate

15. P(can_fly(X) | pelican(X)) = (from 2 and 7)
    P(can_fly(X) | pelican(X) $\wedge$ bird(X) $\wedge$ $\neg$penguin(X)) = (by the independence assumption)
    P(can_fly(X) | bird(X) $\wedge$ $\neg$penguin(X)) > 0.9.

The final conclusion would then be that P(can_fly(tweety) | $\alpha \wedge \beta$) > 0.45.

Note that there are other independence assumptions that are consistent and that lead to opposing conditions. For example, we could make the assumption that can_fly(tweety) is independent of α given bird(tweety). This, together with IND.1, leads directly to the conclusion that P(can_fly(tweety) | $\alpha \wedge \beta$) = 0.9. However, this would conflict with the independence assumption IND.2, and the grounds for the latter are stronger.

The value of P(can_fly(polly) | $\alpha \wedge \beta$) is easily calculated using independence assumptions analogous to IND.1 and IND.2, above.

16. P(can_fly(polly) | $\alpha \wedge \beta$) = P(can_fly(polly) | β) (independence assumption) =
    P(can_fly(polly) $\wedge$ bird(polly) | β) + P(can_fly(polly) $\wedge$ $\neg$bird(polly) | β) =
    P(can_fly(polly) | bird(polly) $\wedge$ β) (from (6)) =
    P(can_fly(polly) | bird(polly)) (independence assumption) = 0.9

Thus, this analysis gives us the desired answers.

Overall, this probabilistic analysis must be considered a partial success. There are assignments of probabilities that are consistent with the given information, and that give the results we expect. Moreover, these assignments can be found by making intuitively justifiable independence assumptions. However, we have only informal criteria for choosing between conflicting independence assumptions. The full automation of this kind of probabilistic reasoning would require a complete formal theory of how independence assumptions are chosen in different cases. Little is known about the structure of such a theory [Pearl, 88].

2.1 Back to the Circuit

We can now construct an analogous probabilistic analysis of our circuit problem. Considering the two situations in figure 4, we would like to deduce that in circuit B, the idealizations are almost certainly valid; while in circuit A, the chances are fifty-fifty whether or not they are valid, depending on whether or not the ball hits the switch.

We give an ad-hoc formalization of these inferences below. (The probabilistic issues would be essentially the same in a more reasonable representation.) Analogously to our construction in the previous section, we will assume that the one-half probability that the switch in circuit A is hit by the ball derives from some observation α, and the certainty that the switch in circuit B is not hit derives from some observation β. The high probability of the idealization holding is measured relative to a state of knowledge not including α or β.

1. circA = series(batA, swA).
   (Circuit A consists of battery A and switch A connected in series.)

2. circB = series(batB, swB).
   (Circuit A consists of battery B and switch B connected in series.)

3. battery(batA) $\wedge$ switch(swA).
   (Categories of the components in A.)

4. battery(batB) $\wedge$ switch(swB).
   (Categories of the components in B.)

5. $\forall_{B,SW}$ [battery(B) $\wedge$ switch(SW) $\wedge$ closed(SW)] $\Rightarrow \neg$ideal(series(B, SW)).
   (A circuit consisting of a closed switch in series with a battery cannot satisfy the idealizations.)

6. $\forall_{C1,C2}$ circuit(series($C1, C2$))
   (The series connection of components forms a circuit.)

7. $\forall_{SW}$ switch(SW) $\Rightarrow$ [closed(SW) $\Leftrightarrow$ was_hit(SW))]
   (A switch is closed just if it has been hit by a falling object.)

8. P(ideal(C) | circuit(C)) = 0.9
   (Circuits can be assumed to obey the idealization.)

9. P(was_hit(swA) | α) = 0.5.
   (There is a one-half chance that switch A was hit.)

10. P(was_hit(swB) | β) = 0.0.
    (There is no chance that switch B was hit.)

We can calculate P(ideal(circA) | $\alpha \wedge \beta$) and P(ideal(circB) | $\alpha \wedge \beta$) as follows.

First, we assume that the event ideal(circA) is independent of event β, which categorizes circuit B. This is analogous to assumption IND.1 of the previous section. We can then deduce

11. P(ideal(circA) | $\alpha \wedge \beta$) = P(ideal(circA) | α) =
    P(ideal(circA) $\wedge$ was_hit(swA) | α) + P(ideal(circA) $\wedge \neg$was_hit(swA) | α)

By (1), (3), (5), and (7) the event ideal(circA) is impossible, given was_hit(swA), so

12. P(ideal(circA) $\wedge$ was_hit(swA) | α) =0.

Proceeding, we can compute

13. P(ideal(circA) $\wedge \neg$was_hit(swA) | α) =
    P(ideal(circA) | $\neg$was_hit(swA) $\wedge \alpha$) $\cdot$ P($\neg$was_hit(swA) | α)

The second factor above, P($\neg$was_hit(swA) | α) = 0.5, by (9).

To compute P(ideal(circA) | $\neg$was_hit(swA) $\wedge \alpha$), we first assume, analogously to IND.2 of the previous section, that ideal(circA) is independent of α given $\neg$was_hit(swA). We then have

14. P(ideal(circA) | $\neg$was_hit(swA) $\wedge \alpha$) = P(ideal(circA) | $\neg$was_hit(swA)).

The event circuit(circA) is certain, given (1) and (6), so

15. P(ideal(circA) | $\neg$was_hit(swA)) = P(ideal(circA) | circuit(circA) $\wedge \neg$was_hit(swA)).

Now, assuming that ideal(circA) is independent of ¬was_hit(swA) given circuit(circA), we get

15. P(ideal(circA) | circuit(circA) ∧ ¬was_hit(swA)) = P(ideal(circA) | circuit(circA)) = 0.9.

Combining (11-15) gives us the desired result P(ideal(circA) | $\alpha \wedge \beta$) = 0.45. The result P(ideal(circB) | $\alpha \wedge \beta$) = 0.9 can be derived analogously.

But the last independence assumption, that ideal(circA) is independent of ¬was_hit(swA) given circuit(circA), is most peculiar. Quite the contrary: Once one knows that the ball did not hit the switch, it is essentially certain that the idealization holds. (As always, of course, there is an open-ended collection of qualification conditions: the components are not broken, we are not in the presence of an immensely powerful contravening electric field, etc. But we are keeping these fixed throughout the entire discussion.) It would seem, in fact, that we could reasonably replace rule (5) above with the stronger statement that a battery and a switch violate idealizations if and only if the switch is closed.

5.a. $\forall_{B,SW}$ [battery(B) ∧ switch(SW)] ⇒ [closed(SW) ⇔ ¬ideal(series(B, SW))].

With this new theory, we would have the equation

15.a. P(ideal(circA) | ¬was_hit(swA)) = 1.0

so P(ideal(circA) | α) = 0.5.

Note, however, that formula (9) "P(ideal(C) | circuit(C)) = 0.9" is now not used anywhere in the derivation. Worse, if we consider a more usual circuit, such as a battery in series with a resistor, the analogous axiom would state that it always satisfies the idealization.

16. $\forall_{B,SW}$ battery(B) ∧ resistor(R) ⇒ ¬ideal(series(B, R)).

But this directly contradicts formula (9)

In fact, to make reasonable sense of the probability (9), we must interpret it relative to a theory in which the structure of the circuits, expressed in (1) and (2) above, is not known. Thus, before you knows the structure of a circuit you may take it as satisfying the idealization with high probability. Once you do know the structure, in most cases you will either be sure that it is satisfied or that it is not.

In a wider context, physical idealizations, like virtually all inferences in this uncertain world, are unquestionably defeasible rather than deductive inferences.

A. If the physical system is only partially specified, it may be a reasonable guess that the idealizations hold. For example, if you see a closed switch in an unknown circuit, it is reasonable guess that the voltage drop across it is negligible.

B. Idealizations are always subject to qualification conditions, which may fail: the components may fail, there may be external interference, and so on.

C. The idealization is wrong but not inconsistent. If so, the idealization is a defeasible inference that may be corrected by additional information, such as a contradictory measurement.

3 Deductive Analysis

We observed in the previous section that, in most cases where the structure of the physical system is known, it is fully determined whether or not it obeys the idealization. It is therefore reasonable to expect that this class of inference should be characterized by a deductive rather than a defeasible theory.

Put it another way: The pelican/penguin theory is indisputably non-monotonic. If we specify that Tweety is a bird, we conclude that it can probably fly; if we further specify that it is a penguin, we withdraw that conclusion. But nothing of the sort occurs with circuits. In most circuits, once the structure is specified, there is no additional information would cause us to decide that they violate the ideal models (again, assuming basic qualification conditions.)[4]

Formulating a deductive theory is a little tricky. The problem is that the theory must state a rule of the form "The idealization holds if and only if Φ," for some condition Φ. The obvious candidate for Φ is the condition that it be consistent that the idealization holds. But that is just the non-monotonic rule that got us into trouble in section 1.

A more careful analysis, however, avoids the problem. The correct form of the rule is "The idealization holds if and only if there is an assignment of voltages and currents consistent with it." The physics will, so to speak, adjust the voltages and currents to fit the idealization, if it can; it will not adjust the position of the switch, or the event of the ball hitting the switch.

We can express this in a deductive logic as follows. We introduce an *assignment* as an ontological sort. An assignment is a labelling of each part of the circuit with a voltage and current. An assignment is ideal if it obeys the idealization. We then state the following rule:

> If there exists an ideal assignment, then the real state of the circuit is equal to that of some ideal assignment. (There may be more than one ideal assignment. For example, in a circuit without a connection to "ground", it is possible to add a constant to the voltage at every node. In a bi-stable circuit, there will be an idealization for each stable state.)

We show in appendix A how this can be axiomatized in a first-order logic. A little care is required to be ensure that an ideal assignment exists if the equations are solvable. To achieve this, we include an axiom that asserts that any numerical labelling of the circuit exists as an assignment, and we give sufficient as well as necessary conditions that an assignment be ideal.

If there are a number of models, of differing degrees of accuracy, or different ranges of applicability, they can combined in a single deductive theory of a form such as the following:

> If there is an assignment satisfying idealization I1, then the real state conforms to I1;
> else if there is an assignment satisfying idealization I2, then the real state conforms to I2;
> ...

This logic works, but its form requires some further discussion. *Why* may an ideal assignment specify voltages and currents but not the position of the switch or the original state of the falling ball? More importantly, when we switch to the next physical domain, how will we know which parameters form part of the "ideal assignment"? Can we be sure that this framework of solution will work at all?

[4] The same considerations apply to the theory of solid object dynamics discussed in [Sandewall, 89]; since the inference structure is monotonic, it is unwise to resort to a non-monotonic theory to describe it. I have shown in [Davis, 90] how that domain can be axiomatized in a deductive theory.

As far as I can tell, the choice of parameters to place in the ideal assignment is a function of the direction of causality. The position of the ball causes the state of the switch, which causes the voltages and currents; the causality does not run in the opposite direction. The direction of the causality from ball to switch is an instance of the general rule that causality does not go backwards in time. The causal arrow from the switch from voltages and currents is not a general phenomenon; in fact, it is just an aspect of the idealization. In an electric bell, there is a causal link from the voltage and current to the acceleration of the clapper [de Kleer and Brown, 83]; in a voltage meter, there is a causal link from the voltage to the position of the needle.

The simplest and most common case of causal direction is the flow of time. In a time-varying system, we suppose that there is an ideal assignment of parameters at each instant of time, and that values calculated in past times may serve as reference parameters for later times, but not vice-versa.

Irreversible causality that is not associated with the time arrow, such as the causal flow from the switch to the electric state, is harder to characterize in a general theory. If there are several levels of irreversible causality, each with an associated set of idealizations, then there must be an "ideal assignment" for each level. If there is a causal connection from level L1 to level L2, then the logic specifies that the ideal assignment for L2 is defined relative to some real state of L1. We can contrive an example of such a hierarchy in our domain, by supposing that the needles of voltage meters in one circuit set the position of the switches in another circuits. (Figure 5). Intuitively, it is easy enough to see what the deductive theory of such a domain should look like; the characteristics of circuit A are independent of the rest; circuit B depends on A; circuits C and D depends on B. (If there is a circularity — the voltage meter setting a switch in the same circuit — then the static theory will generally be inconsistent, and it may be necessary to move to a dynamic theory.) In practice, it would be a painful job to extend the first-order theory of appendix A to handle these kinds of cases. The first task of such a theory would be to identify the separate circuits and their connections, which would require set theory — the whole thing would be a mess. It might be worthwhile defining a special purpose language to describe such a domain.

I am by no means confident that theories of this structure can avoid all anomalies. We shall see (section 5) that dynamic anomalies cannot easily be made to conform to this pattern. There could also be static systems where such theories gave the wrong answer, such as a bi-stable system in which one state observed the idealizations and the other violated them. I have not been able to construct any such examples, but I have no arguments to show that they cannot happen. Ultimately, I am afraid, the only entirely reliable rule is, "The idealization holds if the every truly possible state is close to a state satisfying the idealization, and vice versa," but such a rule is of very little use.

The casual use above of the problematic concept of "causality" deserves some consideration. Is this "causality" the same as the other types of physical causality discussed in the literature? If not, what is it? As far as I can tell, it is not the same as the causality debated between [de Kleer and Brown, 85] and [Iwasaki and Simon, 86], which operates within what we have considered a single causal level. Nor does it seem to correspond to the theories of causality of [Shoham, 88] or [Pearl, 88], since these center around theories of plausible inference, while we are working in a deductive theory.

4 Implications for Implementation

It may be wondered what relevance all this has to any actual implementation of a reasoning system. Consider the ball possibly falling on the switch. The natural implementation of a physical reasoning system would do exactly the right thing with this; it would determine that the ball might or might not hit the switch, consider each case separately, and find the resultant behavior. It would seem rather baroque for the system to reason that, since a closed switch would violate the default, it

should eliminate the possibility that the ball hit the switch. In fact, the only physical prediction system that I know of that reasons backwards in time in that way is, ironically, my own program CHEPACHET (see [Davis, 89] p.427)[5]

However, it seems likely that, as physical reasoning systems grow more sophisticated, temporally or causally backwards reasoning will become increasingly important. For example, the reasoner may be given a collection of measurements taken at various times, and be required to interpolate the behavior of the system in between [Forbus, 86]. In the next section, we will study examples where idealizations impose constraints on physical behavior over extended time; there, again, it seems likely that purely forward reasoning will not suffice. (The same is true of the frame problem. The reason that the frame problem has not yet been problematic in many implemented systems [McDermott, 87] is that these have considered only limited types of inferences.)

5 Idealizations over time

So far, we have considered violations of idealizations that can arise in a single, static situation. There are also idealizations that contradictory when applied over extended periods of time, but are consistent in each situation considered separately. We consider four such examples:

1. Consider a ball rolling on a surface (figure 6). An idealized model, even one that takes contact friction into account, will predict that the ball can roll forever. However, one knows that it will not. The idealized analysis is perfectly acceptable at any single instant, but it leads to contradictory results, applied over all time.

2. Figure 7 shows a circuit containing a battery, an inductor, a resistor, and a switch. The inductor is characterized by the idealization $V = L \cdot dI/dt$, the voltage across the inductor is proportional to the derivative of the current. (L is a constant.) If the switch is open, then there is no current, and all idealizations are satisfied. If the switch is closed, then the current asymptotically approaches the voltage of the battery divided by the resistance of the resistor, and, again, all idealizations are satisfied. However, if the switch is originally closed and then opened, it is impossible for the idealizations to be satisfied. The ideal model of the switch would require that the current drop instantaneously from a finite value to zero, which would make the voltage across the inductor either infinite or undefined. (In reality, opening a switch in series with a powerful electromagnet can create an arc of current across the open switch — very dangerous.)

3. Suppose a ball is bouncing on the floor (Figure 8). The ideal model of partially elastic collisions specifies that the collision reduces the speed of the ball by a fixed fraction μ (the coefficient of restitution). On this assumption, the time between bounces must likewise decrease by a factor of μ each time. Thus, the time between bounces decreases in a geometric series, so the clock times of the bounces converge to a finite time, at which point the ball will be at rest, having just undergone an infinite number of collisions. There is no inconsistency here within Newtonian mechanics — note that both the position and the velocity of the ball are continuous at the cluster point – but there is certainly something anomalous about it. It contradicts an often-discussed principle (Axiom 9 of [McDermott, 82]) that there should not be infinitely many changes of state, or an infinite sequence of separate events, in finite time.[6] We can turn this into a more direct contradiction by supposing that the ball is bouncing on a touch-sensitive toggle switch, which is idealized by the description that whenever it is hit, it turns a light on or off. Now, the problem is clear: is the light on or off when the ball stops bouncing? One of the idealizations has got to give.

[5] Weld's (1988) program EXAG, which makes similar predictions, avoids such reasoning by using a somewhat more complex set of rules.

[6] The status and desirability of this rule will be discussed in a forthcoming paper [Davis, in prep.]

4. Consider the control system illustrated pictorially in figure 9.A and schematically in figure 9.B [Williams, 89]. The liquid is slowly being depleted from the small container by an external process. The aim of the control system is to maintain a roughly constant height of liquid in the container by controlling the flow into the container from a tank. (Historically, this system is a Hellinistic design for maintaining the height of oil in a lamp.) It works by linking the control of the valve to the height of the liquid in the container. When the level in the container is too low, the valve opens, allowing liquid to flow from the tank; when it is too high, the valve closes, and the liquid in the container is gradually depleted by the external process.

Suppose that the valve is modelled by the following idealization:[7]

$$\text{The valve is } \begin{cases} \text{fully open} & \text{if the liquid level is too low.} \\ \text{fully closed} & \text{if the liquid level is too high.} \\ \text{either fully open or fully closed} & \text{if the liquid level is correct.} \end{cases}$$

We assume that the external process creates a constant outflow O from the tank; that when the valve is open, there is an constant inflow $I > O$; and that when it is closed, there is no inflow. Clearly, when the liquid is not at the right level, the system behaves as desired; it fills up if the liquid is too low, and empties if the liquid is too low. What, however, is the behavior of the system once the correct level has been reached? The answer is that there is *no* consistent behavior. In effect, we are trying to solve an equation of the form

$$\dot{x} = \begin{cases} I - O & \text{if } x < x0 \\ -O & \text{if } x > x0 \\ \text{either } I - O \text{ or } -O & \text{if } x = x0 \end{cases}$$

If $x(t0) = x0$, then there is no behavior for $t > t0$ consistent with these equations.[8] If $\dot{x} = I - O > 0$ then x will immediately become greater than $x0$, and $\dot{x}$ will become $-O$. If $\dot{x} = -O$ then x will immediately become less than $x0$ and $\dot{x}$ will become $I - O$.

There are a couple of small modifications to the model of the valve that solve the problem. One possibility is to change the characterization of the valve at the cutoff level. If we specify, instead, that the at the critical control height valve may be anywhere between fully closed and fully open, and thus allows flow anywhere between zero and I, then the equations have the unique solution that, once the height x0 has been reached, the flow through the valve has the constant value O. But, though this solves our current difficulty, it may be a less realistic model of the valve than our original solution. Many valves, including the one shown in figure 9.A, are basically bi-stable. In the steady state they are either fully open or fully closed; they do not maintain a partially open state except for brief transitional periods. With such a valve, the actual physical behavior of the system would be to oscillate between being open and closed, where the period of the transition would depend either on the time necessary to open and close the valve or on the inertia of the liquid. (Liquid takes time to accelerate. Thus, even if the valve opens instantaneously, the liquid does not attain its full level of flow immediately.)

Another solution would be suppose that there is a range of heights in which the valve may be fully open or fully closed. Again, the idealization solves the problem but creates new problems.

[7] For the sake of brevity, my description of the valve below incorporates the functionality of the linkage and of the tank, and thus violates, "No function in structure,"[De Kleer and Brown, 85]. It should be clear, though, that a valid structural description would encounter the same problem.

[8] The reader may wonder where the time asymmetry came from; after all, the effect of time reversal is just to change the sign. The answer is that, in dealing with discontinuous differential equations, it is reasonable to take the derivative of x(t) at t0 to be the limit of x(t1) − x(t0) / t1−t0 as t1 approaches t0 from the right, but not from the left.

It is clear that the techniques discussed in previous sections will not suffice solve these problems. This is particularly evident in case 2 of the bouncing ball; in order to avoid the anomaly at the cluster point, we must assume that the bouncing stops at some previous point. But this is allowing the violation of an idealization in the future to affect the analysis of the past, which is what we have particularly wished to avoid. Cases 1,3 and 4 present similar problems, though in more subtle form. If we wish to solve them using the technique of "ideal assignments" proposed in section 3, then an assignment must cover the state of the system over intervals of time, otherwise it cannot detect the inconsistency. But if so, then it will have the effect of allowing future states to affect earlier states. Such an influence still has all its baneful effects in this extended context. For example, suppose the ball in (2) is bouncing near the edge of a table, and it is undetermined whether or not it will fall off the table before it settles down. We do not want to force the conclusion that it will fall off, in order to avoid the violation of idealization involved in shuddering to a halt on top of the table. But that is exactly the conclusion we will get if we take an "ideal assignment" to be a history of the ball over time.

The point, let me emphasize, is not that we don't have models that can avoid these anomalies. We do. But we would prefer to use the simpler models if possible. We are looking for a theory that will allow us to use a simple model, when it works, without computing the results of a more complex model. We found such a theory for the simple circuit of sections 1-3; but the techniques used there will not work with the examples in this section.

6 Methodological Objections

It can be argued, from a number of different points of view, that our difficulties come about purely because we have taken an ill-chosen view of our problem, and applied an unsuitable methodology. Let me present and contest three such arguments.

The Physicist's Response. "The idealizations you're discussing are just consequences of ignoring lower-order terms in the correct theory, and should be viewed as such. One approach would be to begin by applying the correct theory to the specific problem, and then throw out the lower-order terms. Another approach would be to have rules, that are justifiable in the correct theory, that state, 'Under circumstances X,Y,Z, idealization I1 will give answers correct within ϵ.' But it is simply foolish to apply idealizations outside their range of applicability."

"Further proof that this is the correct way to view the issue," (the physicist proceeds) "is that a real scientist often uses an idealization at one point in a chain of reasoning and to use a more detailed model at another point. For example, in calculating the tides, it would be reasonable to treat the earth with its seas as a rigid object when calculating the orbit of the moon, and then to treat the seas as non-rigid in finding the tides. Now, the only way to make sense of this is to suppose that the reasoner has an single overarching theory from beginning to end, and he approximates it in various places by idealizations to simplify his calculations. Certainly if AI researchers want to separate knowledge from control, they cannot view the reasoner as using conflicting sets of axioms in different stages of reasoning."

In many cases, including electronic circuits, this seems incontestable. In many other cases, though, it is pure fantasy. Often, particularly in non-physical idealizations such as economic models, there may not be any more accurate model to go to. Often, even if there is a more accurate model, it is entirely unclear what that model predicts, or under what circumstances it is accurately approximated by the idealization. In reality, scientists and engineers go around all the time using idealizations with only the dimmest notion, or no notion at all, of what lower-order terms they are ignoring or how large they are. No doubt, these experts over time develop a "seat-of-the-pants" feeling for when their idealizations are OK, and no doubt it is ultimately part our task, as builders of AI physical reasoners,

to capture this feeling in computational terms. However, I think that, without undue cynicism, we can suppose that this feeling can often be more accurately approximated as pure ignorance than as a set of sufficient conditions formally justified by a well-established theory.

Furthermore, in many cases, there are a number of lower-order terms that could potentially become important when the idealization fails; determining which of these is in fact dominant may require whole new types of information. In the valve example of section 5, the terms being ignored may include the time required to open the valve, or a range of half-open states, or the inertia of the liquid. Which of these is dominant may depend on such issues as the flexibility of the rod or the viscosity of the liquid which have not been specified. It may be possible to show the disjunction, "Either term 1 is large or term 2 is large or ..." but this surely is derived from the inconsistency of the idealization, rather than from prior estimates of the terms themselves.

Given that a physical reasoner is using an idealized theory without reference, direct or indirect, to a more accurate theory, then he is more or less committed to accepting the theory as long as its predictions make sense. If the results are consistent, what reason does the reasoner have to reject them? (There are, actually, other criteria. For example, one might reject the predictions of an idealization if the parameters are close to values that generate an inconsistency. Or one might reject them if a comparable theory gave incorrect answer on an analogous example. But, surely, such considerations are "closer" to our approach than to the hypothetical physicist's.)

The die-hard Hayesian.[9] "It is nonsense to speak of 'some cases leading to an inconsistency in the model.' It is a basic part of a commonsense understanding of the domain that it is possible to attach a wire across a battery. Given that fact, the model is simply inconsistent with no qualification. Now, there is no point whatever in trying to make sense of an inconsistent theory. Go find a consistent, commonsense theory of the domain, and then there is no problem."

The answer is that this is precisely what we are trying to do. We can just rephrase, "Cases where the idealization is inconsistent," as "exceptional cases." It would be nice, of course, to find a uniform theory that didn't have to make exceptions, but there's no assurance that such a theory can be found.

The limited-aims advocate. "You're trying to stuff too much into your object-level theory. All that it's worth trying to get from a logical formalization is limited coverage of a microworld. It is perfectly reasonable to have a theory which states that a battery always maintains a constant voltage and that a wire always maintains equal voltage. In real circuit design problems, these will not lead to contradiction. If you go to another context, you can use a different theory. But there is no point in trying to formalize the choice between microworlds at the same level as the microworlds themselves. Choosing an appropriate microworld depends on external criteria: the class of physical situations considered, the type of problem (prediction, design, explanation, etc.) addressed, the accuracy required, the time available, and so on."

This kind of argument is hard to counter. The only real answer to "It's a waste of time to work on problem XYZ" is "Here is a solution to XYZ", which I haven't got. The most that I can say is that the argument seems overly pessimistic. Admittedly, a theory cannot very well express constraints on the time required to use itself or on the direction of inference without horrendous mixings of levels. But there is no reason that an object-level theory cannot express the accuracy of the results of some rule, or the class of physical situations where it applies. In fact, any characterization of these issues should be translatable into purely object-level terms. It is hard to see why it should not be possible to merge them with the object-level theory or theories.

[9] I am not, of course, claiming that this argument would be endorsed by Pat Hayes, or by anyone else other than myself in some moods.

7 Appendix A: Detailed Deductive Theory

This appendix presents the first-order theory sketched in section 3. The electronic circuitry is axiomatized at the level of standard basic circuit analysis. The axiomatization of the falling ball and its interaction with the switch is much more *ad hoc,* but it should be evident that it can be replaced by a more realistic theory without affecting basic structure or inferential power of the theory.

We use a sorted first-order theory with equality. The temporal logic follows [McDermott, 82] in taking time to be real-valued; however, we do not need branching time lines.

We use the following sorts of entities:

- *Objects.* The objects we use are balls and electronic components, such as batteries and switches.
- *Nodes.* A node is a place where two or more components are connected.
- *Circuits.* A circuit is a collection of components and nodes.
- *Voltages* form a real-valued, integral quantity space.
- *Voltage differences* form the difference space of voltages.
- *Currents* form a real-valued, differential quantity space.
- *Resistances* form a real-valued, differential quantity space, the quotient space of voltage difference divided by current.
- A *place* is a location in geometric space.
- A *situation* is an instant in the time line.
- An *assignment* is an association of a voltage with every node, and of a current with every pair of a component and a node.
- A *fluent* is a function from situations and assignments to some other space. In our theory, we will use fluents that range over voltages and currents.

The sorts of variables will be indicated by the first letter of the variable name, as indicated in Table 1.

Letter	Sort
A	Assignment
C	Circuit
F	Fluent
I	Current
N	Node
O	Object
S	Situation
V	Voltage
W	Situation or assignment
X, Y	Object or Node
Z	Arbitrary

Table 1: Key of Sorts by Variable Name

Table 2 shows the non-logical primitives, aside from arithmetic primitives such as $+$. Table 3 shows the physical axioms. Table 4 shows the boundary conditions of our "ball hitting switch" example.

Equality:

distinct($Z1 \ldots Zk$) — Predicate. Entities $Z1 \ldots Zk$ are all distinct.

Temporal primitives:

true_in(W, F) — Predicate. Truth-valued fluent (state) W is true in situation or assignment W.

value_in(W, F) — Function. The value of fluent F in situation or assignment W.

$S1 < S2$ — Predicate. Situation $S1$ precedes situation $S2$.

Assignments:

ideal(A, S, C) — Predicate. Assignment A is a ideal assignment for circuit C in situation S.

agrees(A, S, C) — Predicate. Assignment A gives the true voltages and currents for circuit C in situation S.

Electronics:

battery(O) — Predicate. Object O is a battery.

switch(O) — Predicate. Object O is a switch.

resistor(O) — Predicate. Object O is a resistor.

component(O, C) — Predicate. Object O is a component of circuit C.

node(N, C) — Predicate. Node N is a node of circuit C.

joins1(X, Y) — Predicate. X and Y are connected. Either X is a component and Y is a node or vice versa.

joins($X, Y1 \ldots Yk$) — Predicate. $Y1 \ldots Yk$ is a list, without repetition, of the things connected to X. Either X is a component and $Y1 \ldots Yk$ are nodes, or vice versa.

volts_of(O) — Function. Voltage rating of battery O (a voltage difference).

resistance_of(O) — Function. Resistance of resistor O.

voltage_of(N) — Function. The fluent giving the voltage at node N in each situation.

current_into(X, Y) — Function. The fluent giving the current into X from Y in each situation. Either X is a component and Y is a node or vice versa.

place(O) — Function. The fluent giving the place of object O in each situation.

closed(O) — Function. The fluent of switch O being closed.

ideal_comp(A, O, S) — Predicate. Assignment A satisfies the idealization of component O in situation S.

kcl(A, X) — Predicate. Assignment A satisfied Kirchoff's Current Law at component or node X.

Other Physical:

ball(O) — Predicate. Object O is a ball.

unsupported(O) — Function. The fluent of ball O being unsupported.

above($R1, R2$) — Predicate. Place $R1$ is above place $R2$.

Table 2: Non-logical primitives

Assignments:

1. $\forall_{S,C}\ [\exists_{A1}\ \text{ideal}(A1,S,C)] \Rightarrow \exists_{A2}\ \text{ideal}(A2,S,C) \wedge \text{agrees}(A2,S,C)$.
   (Rule of idealizations: If situation S has some ideal assignment $A1$ on circuit C, then the actual state of C in S agrees with some ideal assignment $A2$.)

2. $\text{ideal}(A,S,C) \Leftrightarrow$
   $[[\forall_O\ \text{component}(O,C) \Rightarrow \text{ideal_comp}(O,A,S) \wedge \text{kcl}(A,O)] \wedge$
   $[\forall_N\ \text{node}(N,C) \Rightarrow \text{kcl}(A,N)]]$.
   (An assignment is ideal just if it observes the idealized model of every component and observes Kirchoff's current law at every component and node. Kirchoff's voltage law is automatically insured by the device of associating voltages with nodes.)

3. $\forall_{N1\ldots Nm}\ \forall_{O1\ldots Ok}\ \forall_{I11\ldots Imk}\ \forall_{V1\ldots Vm}$
   $[\text{distinct}(O1\ldots Ok) \wedge \text{distinct}(N1\ldots Nm)] \Rightarrow$
   $\exists_A\ \text{value_in}(A,\text{voltage}(Nj)) = Vj\ (j = 1\ldots m) \wedge$
   $\text{value_in}(A,\text{current_into}(Oi,Nj)) = Iij\ (i = 1\ldots k, j = 1\ldots m)$.
   (Axiom schema for integer k,m: Any association of voltages with nodes and currents with nodes and components corresponds to some assignment.)

4. $\text{agrees}(A,S,C) \Leftrightarrow$
   $[\forall_{N,O}\ [\text{node}(N,C) \wedge \text{component}(O,C)] \Rightarrow$
   $[\text{value_in}(A,\text{voltage}(N)) = \text{value_in}(S,\text{voltage}(N)) \wedge$
   $\text{value_in}(A,\text{current_into}(O,N)) = \text{value_in}(S,\text{current_into}(O,N))]]$
   (An assignment A agrees with a situation S on circuit C if they assign the same voltages and currents to the nodes and components of C.)

Electronics;

5. $\text{kcl}(A,X) \Leftrightarrow$
   $[\forall_{Y1\ldots Yk}\ \text{joins}(X,Y1\ldots Yk) \Rightarrow$
   $\text{value_in}(A,\text{current_into}(X,Y1)) + \ldots + \text{value_in}(A,\text{current_into}(X,Yk)) = 0]$
   (KCL: An assignment A obeys KCL on node or component X if the currents into X from the attached components or nodes sum to zero. Axiom schema for integer k.)

6. $[\text{battery}(O) \wedge \text{joins}(O,N1,N2)] \Rightarrow$
   $[\ \text{ideal_comp}(A,O,S) \Leftrightarrow$
   $\text{value_in}(A,\ \text{voltage_of}(N2)) - \text{value_in}(A,\ \text{voltage_of}(N1)) = \text{volts_of}(O) > 0.]$
   (Ideal battery characteristic: The voltage drop across a battery is equal to the rating of the battery.)

7. $[\text{switch}(O) \wedge \text{true_in}(S,\text{closed}(O)) \wedge \text{joins}(O,N1,N2)] \Rightarrow$
   $[\ \text{ideal_comp}(A,O,S) \Leftrightarrow$
   $\text{value_in}(A,\text{voltage_of}(N1)) = \text{value_in}(A,\text{voltage_of}(N2))]$.
   (Ideal closed switch characteristic: There is no voltage drop across a closed switch.)

8. $[\text{switch}(O) \wedge \neg\text{true_in}(S,\text{closed}(O)) \wedge \text{joins}(O,N1,N2)\ ] \Rightarrow$
   $[\ \text{ideal_comp}(A,O,S) \Leftrightarrow$
   $\text{value_in}(A,\text{current_into}(O,N1)) = \text{value_in}(A,\text{current_into}(O,N2)) = 0.]$
   (Ideal open switch characteristic: There is no current into a closed switch.)

9. [resistor(O) $\wedge$ joins($O, N1, N2$)] $\Rightarrow$
[ideal_comp(A, O, S) $\Leftrightarrow$
value_in(A,current_into($N1, O$) $\cdot$ resistance(O) =
(value_in(A,voltage_of($N2$)) $-$ value_in(A,voltage_of($N1$)). (Ideal resistor characteristic: The voltage drop across a resistor is equal to the current times the resistance.)

10. value_in(A,current_into(O, N)) $= -$value_in(A,current_into(N, O)).
(The current from a node to a component is the negative of the current from a component to a node. Note that this does not contradict (3), since (3) only allows one to specify the currents into objects, not to independently specify the currents into nodes. The purpose of this convention is just to simplify the statement of KCL (5).

Circuit description coherence:

11. joins($X, Y1 \ldots Yk$) $\Leftrightarrow$
[$Yi \neq Yj$ $(i, j = 1 \ldots k, i \neq j)$ $\wedge$
[$\forall_Y$ joins1(X, Y) $\Leftrightarrow$ [$Y = Y1 \vee \ldots Y = Yk$]]].
(Coherence of joining statements: X joins $Y1 \ldots Yk$ if $Y1 \ldots Yk$ are a list without repetitions of the elements joined individually to X. Axiom schema for integer k.)

12. joins1(X, Y) $\Rightarrow$ joins1(Y, X).
(Individual joining is a symmetric relation.)

13. joins1(O, N) $\Rightarrow$ [component(O, C) $\Leftrightarrow$ node(N, C)].
(If O is joined to N, then they are part of the same circuit. Intuitively, a circuit is a minimal set of nodes and components satisfying this definition, but this is hard to express in a first-order theory.)

Falling balls

14. [$\neg$true_in($S1$,closed(OS)) $\wedge$ switch(OS)] $\Rightarrow$
[[$\exists_{S2}$ $S2 > S1 \wedge$ true_in($S2$,closed(OS))] $\Leftrightarrow$
$\exists_{OB}$ ball(OB) $\wedge$ true_in($S1$,unsupported(OB)) $\wedge$
above(value_in($S1$,place(OB)), value_in($S1$,place(OS))]. (An open switch eventually closes just if there is an unsupported ball above the switch.)

Distinct

15. distinct($Z1 \ldots Zk$) $\Leftrightarrow$ $Zi \neq Zj$ $(i, j = 1 \ldots k, i \neq j)$
(Definition of distinct.)

Table 3: Axiomatization

P1. battery(obat) $\wedge$ switch(osw) $\wedge$ ball(oball).
(We have a battery, a switch, and a ball.)

P2. joins(obat, n1, n2) $\wedge$ joins(osw, n2, n1).
(The switch is connected across the battery.)

P3. obat $\neq$ osw.
(The battery is not the switch.)

P4. $\forall_O$ component(O,circuit1) $\Leftrightarrow$ [O =obat $\vee$ O =osw].
(The components of the circuit are the battery and the switch.

P5. true_in(s1, unsupported(oball)).
(The ball is unsupported in s1.)

P6. $\neg$true_in(s1, closed(osw)).
(The switch is open in s1.)

P7. $\forall_O$ above(value_in(s1, place(O)), value_in(s1, place(osw))) $\Rightarrow$ O =oball.
(No objects other than the ball are above the switch.)

HA. above(value_in(s1, place(oball)), value_in(s1, place(osw))).
(Hypothesis A: The ball is above the switch in s1.)

HB. $\neg$above(value_in(s1, place(oball)), value_in(s1, place(osw))).
(Hypothesis B: The ball is above the switch in s1.)

CA. $\neg[\forall_{S2}\ S2 >$s1 $\Rightarrow \exists_A$ ideal($A, S2$,circuit1)]
(Conclusion A: At some future time, the circuit will violate the idealization.)

CB. $\forall_{S2}\ S2 >$s1 $\Rightarrow \exists_A$ ideal($A, S2$,circuit1)
(Conclusion B: The circuit will always respect the idealization.)

Table 3: Problem Statement

The following inferences can now be demonstrated.

- Conclusion CA is a consequence of the theory consisting of axioms (1-15), constraints (P1-P7), and hypothesis HA, but conclusion CB is not.
- Conclusion CB is a consequence of the theory consisting of axioms (1-15), constraints (P1-P7), and hypothesis HB, but conclusion CA is not.
- Neither CA nor CB is a consequence of the theory consisting just of axioms (1-15) and constraints (P1-P7).

8 References

Addanki, Sanjaya, Roberto Cremonini, and J. Scott Penberthy (1989). "Reasoning about Assumptions in Graphs in Models." In D. Weld and J. de Kleer (eds.) *Qualitative Reasoning about Physical Systems,* Morgan Kaufmann, San Mateo, CA.

Davis, Ernest (1989). "Order of Magnitude Reasoning in Qualitative Differential Equations," In D. Weld and J. de Kleer (eds.) *Qualitative Reasoning about Physical Systems,* Morgan Kaufmann, San Mateo, CA.

Davis, Ernest (1990). *Representations of Commonsense Knowledge,* Morgan Kaufmann, San Mateo, CA.

Davis, Ernest (in preparation) "Zeno's Revenge: Infinite Loops in Finite Time."

de Kleer, Johan and John Seely Brown (1983). "Assumptions and Ambiguities in Mechanistic Mental Models," in D. Gentner and A. Stevens (eds) *Mental Models,* Lawrence Erlbaum Assocs., Hillsdale, NJ.

de Kleer, Johan and John Seely Brown, (1985). "A Qualitative Physics based on Confluences," In Daniel Bobrow (ed.) *Qualitative Reasoning about Physical Systems,* MIT Press.

Forbus, Kenneth (1986). "Interpreting Measurements of Physical Systems," *AAAI-86* pp. 113-117.

Hanks, Steven and Drew McDermott, (1987). "Nonmonotonic Logic and Temporal Projection," *Artificial Intelligence,* vol. 33, pp. 379-412.

Iwasaki, Yumi and Herbert Simon, (1986). "Causality in Device Behavior," *Artificial Intelligence,* vol. 29, pp. 3-32.

McCarthy, John, (1980). "Circumscription — A Form of Nonmonotonic Logic." *Artificial Intelligence,* vol. 13, pp. 27-39.

McDermott, Drew and Jon Doyle, (1980). "Non-Monotonic Logic I," *Artificial Intelligence,* vol. 13, pp. 41-72.

McDermott, Drew (1982). "Nonmonotonic Logic II: Nonmonotonic Modal Theories," *JACM,* vol. 29, pp. 33-57.

McDermott, Drew (1987). "AI, Logic, and the Frame Problem." in Frank Brown (ed.) *The Frame Problem in Artificial Intelligence,* Morgan Kaufmann.

Moore, Robert (1985). "Semantical Considerations on Nonmonotonic Logic," *Artificial Intelligence,* vol. 25, pp. 75-94.

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,* Morgan Kaufmann, San Mateo, CA.

Raiman, Olivier, (1986). "Order of Magnitude Reasoning." *AAAI-86,* pp. 100-104.

Reiter, Ray, (1980). "A Logic for Default Reasoning." *Artificial Intelligence,* vol. 13, pp. 81-132.

Sandewall, Erik (1989) "Combining Logic and Differential Equations for Describing Real-World Systems." In R. Brachman, H. Levesque, and R. Reiter (eds.) *Proc. First Intl. Conf. on Principles of Knowledge Representation,* Morgan Kaufmann, San Mateo, CA.

Stein, Lynn and Leora Morgenstern, (1989) "Motivated Action Theory: Formal Theories of Action." Brown University Tech. Rep.

Shoham, Yoav (1988) *Reasoning about Change: Time and Causation from the Standpoint of Artificial Intelligence,* MIT Press.

Weld, Daniel (1988). "Exaggeration," *AAAI-88*, pp. 291-295.

Williams, Brian (1989). "Invention from First Principles via Topologies of Interaction." Ph.D. thesis, MIT.