The Naive Physics Perplex

Ernest Davis*
Courant Institute
New York University

July 23, 1997

Abstract

The "Naive Physics Manifesto" of Pat Hayes [1978] proposes a large-scale project of developing a formal theory encompassing the entire knowledge of physics of naive reasoners, expressed in a declarative symbolic form. The theory is organized in clusters of closely interconnected concepts and axioms. More recent work in the representation of commonsense physical knowledge has followed a somewhat different methodology. The goal has been to develop a competence theory powerful enough to justify commonsense physical inferences, and the research is organized in *microworlds*, each microworld covering a small range of physical phenomena. In this paper we compare the advantages and disadvantages of the two approaches. We also discuss some difficult key issues in automating commonsense physical reasoning.

1 Three Scenarios

Consider the following scenario:

> **Scenario 1:**
> A gardener who has a valuable plant with a long delicate stem protects it against the wind by *staking* it; that is, by plunging a stake into the ground near the plant and attaching it to the stake with string. (Figure 1.)

We might not all manage to think up this contrivance, faced with this problem, but we can all understand how it works. This understanding is manifested in a number of different abilities:

- We can give an *explanation* of the problem and the solution. That is, we can generate a text along the following lines: "The wind may bend the plant; the fragile stem, bent too far, may snap, killing the plant. But if the plant is staked, then the string holds it in place, preventing any extreme bending. The string, in turn, is held in place by the stake, which, being comparatively stiff, is not bent either by the wind or by the force of the wind against the plant as transmitted through the string, and, being stuck in the ground, remains upright."

*This paper originated in talks given at the University of York and the University of Leeds. Thanks to my hosts there, Tony Cohn and Alan Frisch. Thanks also to Leora Morgenstern for encouragement and helpful suggestions. The intellectual debt of this paper to Pat Hayes should be entirely evident. The writing of this paper has been supported by NSF grants #IRI-9300446 and #IRI-9625859.

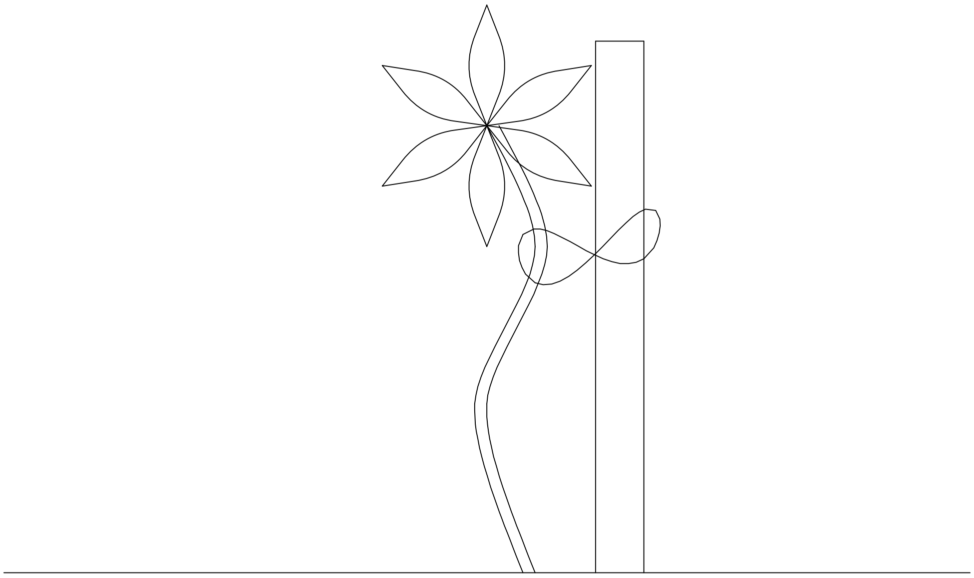


Figure 1: Staking a Plant

- We can *carry out* the plan, which involves both hand-eye coordination and also the reasoning ability to fill in implicit steps of the plan. For example, the string must be looped around the stake and the plant and tied. Since the plan, as given above, does not specify this step, the reasoner must infer it.

- We can *adapt* this solution to other problems, or adapt it to give alternative solutions to this same problem. For example, plants are sometimes staked to prevent their breaking under their own weight. An alternative to staking may be to encircle the plant with a metal frame.

- We can answer questions about *variants* of the plan. What would happen: If the stake is only placed upright on the ground, not stuck into the ground? If the string were attached only to the plant, not to the stake? To the stake, but not to the plant? If the plant is growing out of rock? Or in water? If, instead of string, you use a rubber band? Or a wire twist-tie? Or a light chain? Or a metal ring? Or a cobweb? If, instead of tying the ends of the string, you twist them together? Or glue them? Or place them side by side? If you use a large rock rather than a stake? If the stake is very much shorter than the plant? If the string is very much longer, or very much shorter, than the distance from the stake to the plant? If the distance from the stake to the plant is large as compared to the height of the plant? If the stake is also made out of string? Trees are sometimes blown over in heavy storms; can they be staked against this?

It would seem that the depth and power of our understanding is most readily exhibited by this last-mentioned ability of exploring variants. Over a limited class of plans, explanations and execution sequences can be canned, or generated by very narrow special-purpose techniques. Moreover, the difficulties in writing an adaptable text generator or plan executor are mostly those of natural language and of robotics, respectively; in practice, these issues swamp the problems of representation and reasoning. Adaptation and alternative application of plans certainly shows understanding, but may require a level of ingenuity that is not always reasonable to expect. But anyone with an understanding of the scenario should certainly be able to say something about how things change or stay the same under small changes of the situation or the plan; and, conversely, so many different variations can be hypothesized that intelligent answers can only be attained with some large degree of understanding.

Let us broaden our view by considering two more scenarios, with variants.

> **Scenario 2:** (due to Leora Morgenstern [private communication])
> In baking cookies, once you have the cookie dough prepared, you first lightly spread flour over a large flat surface; then roll out the dough on the surface with a rolling pin; then cut out cookie shapes with a cookie cutter; then put the separated cookies separately onto a cookie dish and bake,

What happens if: You do not flour the surface? You use too much flour? You do not roll out the dough, but cut the cookies from the original mass? You roll out the dough but don't cut it? You cut the dough but don't separate the pieces?

What happens if the surface is covered with sand? Or covered with sandpaper? If the rolling pin has bumps? or cavities? or is square? If the cookie cutter does not fit within the dough? What happens if you use the rolling pin just in the middle of the dough and leave the edges alone? If, rather than roll, you pick up the rolling pin and press it down into the dough in various spots? Ordinarily the cutting part of the cookie cutter is a thin vertical wall above a simple closed curve in the plane; suppose it is not thin? or not vertical? or not closed? or a multiple curve? If the cuts with the cutter overlap?

Does the dough end up thinner or thicker if you exert more force on the rolling pin? If you roll it out more times? If you roll the pin faster or slower? Do you get more or fewer cookies if the dough is rolled thinner? If a larger cookie cutter is used? If there is more dough? If the cuts with the cutter are spread further apart?

> **Scenario 3:**
> The following experiment is described in [Shakhashiri, 85] for estimating absolute zero using household objects. Prepare a pot of boiling water and a pot of ice water. Take a graduated baby bottle and hold it (using tongs) in the boiling water. After a few minutes, when it has stopped bubbling, remove it and plunge it rapidly into the ice water. Water will then stream into the baby bottle through the nipple, as the gas contracts. (Actually, the nipple collapses; to allow the flow of water, you have to manipulate the nipple.) When the flow of water stops, the volume of the water that has entered the bottle may be measured by holding the bottle right-side up; the final volume of the gas at 0° C may be measured by holding the bottle upside down. The initial volume of the gas at 100° C is the sum of the final volume of the gas plus the volume of the water. By doing a linear extrapolation between these two values to the point where the volume of the gas would be zero, one can find the value of absolute zero.[1] (Figure 2).

What would happen: If the bottle is immersed only very briefly in the hot water? Or only very briefly in the cold water? If it is laid on top of the pots of water rather than immersed in them? If the bottle is left in the outside air for a long time between being in the hot water and being in the ice water? If the bottle has an open end with no nipple? If the nipple has no hole? If the bottle has other holes besides this nipple? If the bottle is opaque? If you use containers with air at 100° and 0° rather than water? If the quantity of ice water in the second pot is very small? very large?

[1] I tried this experiment three times. Twice, the entire baby bottle was crushed by the pressure in the cold water. The one time it ran successfully, it gave a value of −300° C for absolute zero, the true value being −273° C — not bad, for a baby-bottle experiment.

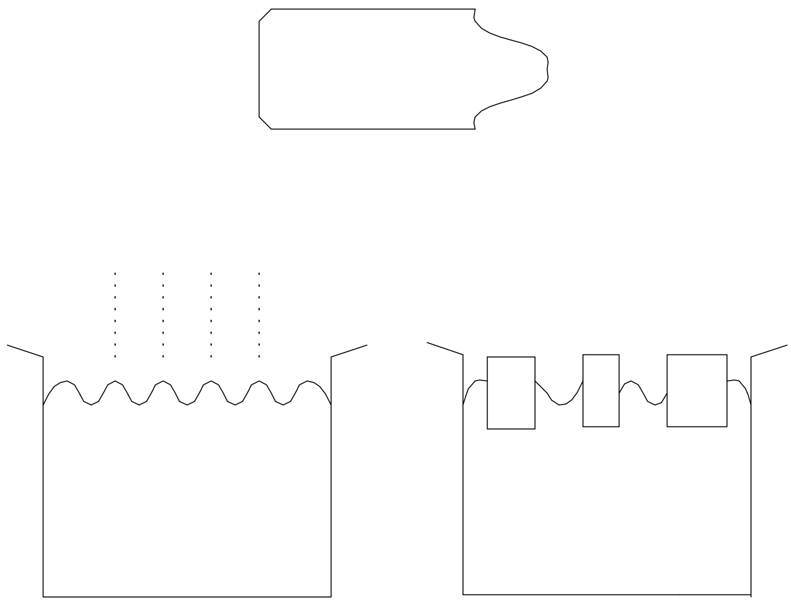


Figure 2: Determing Absolute Zero

or if the quantity of hot water in the first pot is very small or very large? If the bottle is coated with styrofoam? If the bottle is opaque? If the bottle is not graduated? Why is the following not a reasonable experiment: “Take a volume of gas in your hands; cool it; and see how much it shrinks.”

2 Commonsense Physics

These three scenarios above exhibit a number of characteristic properties:

- They rely almost entirely on *commonsense knowledge*; that is, knowledge acquired informally at an early age, rather than explicitly taught. Scenario 3 requires an understanding of the thermal expansion of gasses, which is usually “book learning”. All other aspects of this scenario, and all aspects of scenarios 1 and 2 are commonsensical. A naive subject who has been introduced to thermal expansion should be able to answer almost all the variant questions.

- Quantitative relations are important; recall such questions as, “What happens if the string is much shorter than the distance from the stake to the plant?” or “What happens if the quantity of cold water is very small?” However, precise quantitative values are rare and textbook style equations are practically non-existent, with the exceptions, again, of the values 0°C and 100°C and the linear equation of thermal expansion.

- Similarly, geometric properties and relations are important: The bottle must not have holes other than the nipple and must be immersed in the water. The bowl in scenario 2 must be rightside up. But no precise geometric descriptions are given or needed.

- Each scenario involves a range of types of materials and processes. Scenario 1 involves the somewhat flexible plant, the gaseous wind, the rigid stake, the very flexible string, and the penetrable earth. Scenario 2 involves the malleable cookie dough and the rigid rolling pin, cookie cutter, and surface. Scenario 3 involves the solid baby bottle, the liquid water, and the gaseous air.

- All three scenarios involve the manipulatory powers of an agent. Scenario 3, though not scenarios 1 and 2, also involves perceptual powers. The facts that the experimenter cannot simply cool a volume of gas that he holds in his hands, or that he cannot easily measure quantities in an opaque or ungraduated bottle, must be understood for these alternative experimental designs to be rejected.
- All three scenarios lie outside the range of current automated reasoners. Since I have in the past [Tuttle, 1993] been accused of giving an overly rosy impression of the state of the theory of automated commonsense reasoning, let me stress this point: *As far as I know, no one currently knows how to automate these inferences nor how to represent the knowledge used in them. I do not believe that this will be known any time in the near future.* The purpose of these three example scenarios is to indicate a direction for study and an ultimate goal, not to illustrate the capacities of existing programs or theories.

3 The Naive Physics Manifesto

Commonsense physical reasoning was first and most famously promoted as a domain for AI research by Pat Hayes [1978] in the "Naive Physics Manifesto".[2] That paper advocated a research programme of developing a formalization of naive physics satisfying the following four criteria:

- Thoroughness. "It should cover the whole range of everyday physical phenomena."
- Fidelity. "It should be reasonably detailed."
- Density. "The ratio of facts to concepts should be fairly high."
- Uniformity. "There should be a common formal framework for the whole formalization." Hayes expresses a preference for first-order logic or some extension thereof, but does not insist on it. What is critical, in his view, is that the representation have a clear interpretation.

All considerations of implementation, application, or inference strategy are to be deferred until the formalization is largely complete. "It is not proposed to make a computer program which can 'use' the formalism in some sense. For example, a problem-solving program or a natural language comprehension system with the representation as target. [Such programs] have several ... dangerous effects. It is perilously easy to conclude that because one have a program that *works* (in some sense), its representation of its knowledge must be more or less *correct* (in some sense). Regrettably, the little compromises and simplifications needed in order to get the program to work in a reasonable space or in a reasonable time can often make the representation even less satisfactory that it might have been." Hayes further remarks "The decision to postpone details of implementation ocan be taken as an implicit claim that the representation content of a large formalisation can be separated fairly cleanly from the implementation decision; this is by no means absolutely obvious, although I believe it to be substantially true." This last point, of course, is a central point of attack by such critics as McDermott [1987].

The large theory of naive physics is structured in terms of *clusters*, a cluster being a nexus of concepts tightly related by a rich collection of axioms. Hayes gives the following examples of clusters:

[2] All quotations in this section are taken from [Hayes, 78]. The published version of this is always cited as [Hayes, 79]; however, I have never actually set eyes on this, and I don't know what changes may have been made before publication. The later version [Hayes, 85] is a substantially different paper.

"measuring scales", "shape, orientation, and direction", "inside and outside", "histories", "energy and effort", "assemblies", "support", "substances and physical states," "forces and movements", and "liquids". A large part of the paper is devoted to preliminary analysis of these various clusters. The companion paper "Ontology for Liquids" [Hayes, 85] is an in-depth analysis of the "liquids" cluster.

The question of finding the proper organization into clusters is considered one of the key issues in the enterprise:

> Identifying these clusters [of tightly associated concepts] is both one of the most important and one of the most difficult methodological tasks in developing a naive physics ... The symptom of having got it wrong is that it seems hard to say anything very useful about the concepts one has proposed ... But this can also be because of having chosen one's concepts badly, lack of imagination, or any of several other reasons. It is easier, fortunately, to recognize when one is in a cluster: assertions suggest themselves faster than one can write them down.

(I must confess that I personally have never attained the state of grace described in the last sentence above. In my experience, formalization is always a slow and delicate process, and a great deal of care is needed to avoid inconsistencies, unintended consequences, and gaps.)

Hayes proposes that the research programme be carried out by a committee. Each member of the committee will be assigned a particular cluster to formalize. The committee will meet from time to time in order to integrate their various efforts into a larger theory. This integration will no doubt require that formalizations of clusters be reworked, that new clusters be investigated, and that old ideas for clusters that prove to be useless be discarded.

One issue that Hayes discusses very little, rather curiously, is the choice of naive physics as a domain for study. He does say that "One of the good reasons for choosing naive *physics* to tackle first is that there seems to be a greater measure of interpersonal agreement here than in many fields," but he does not indicate what the other reasons might be. To my mind, the chief other advantages of naive physics as contrasted with, say, folk psychology or naive social science are:

- The power of "real" physics, the paradigm of a theory that is comprehensive, exact, and correct. The metatheoretic, mathematical, and logical structures have been extensively studied. Vast amounts of software carrying one or another type of computation in this domain have been implemented. Of course, naive physics is quite different from real physics; still, this give us an immense body of reliable knowledge on which to draw.
- Problems of intensionality and self-reference do not arise. Physics is a purely extensional theory.
- A broad range of practical applications.

4 Two Common Misconceptions

There are two common misimpressions of Hayes' proposal. The first is an understandable confusion. Seeing that the Naive Physics Manifesto and the Ontology for Liquids are full of formulas written in first-order logic and formal proofs, many people have gotten the false idea that Hayes is proposing that a reasoning program should explicitly manipulate logical formulas using some general purpose theorem-proving method. Now, various people (e.g. [Moore, 1982], [Kowalski, 79]) do indeed advocate this view, but Hayes does not, at least not in these papers.[3] He is, in fact, entirely agnostic as to how the knowledge should be implemented as data structures or what procedures should manipulate it. Hayes' proposal is to analyze naive physical reasoning at the *knowledge level* [Newell, 1980], in terms that are independent of the particular computing architecture, algorithms, and data structure. First-order logic is chosen as a language to describe the knowledge level, precisely because it is a neutral one, that does not presuppose any particular form of implementation.

The intended relation between a logical domain theory and a reasoning program is similar to the relation between a programming language semantics and a compiler. The semantics specifies what the compiler should do; a compiler is correct if the semantics of the output code is compatible with the semantics of the source code. But one does not necessarily expect a compiler to be written in the abstruse formalisms of programming language semanticists. Similarly, the desired relation between a logical domain theory and a reasoning program is that the theory should *characterize* or *justify* the actions of the program, in the sense that some significant part of the results computed by the program corresponds to, or approximates, valid conclusions in the theory. But the internals of the program need not contain anything that looks like the theory.

For example, STRIPS-style planners can be characterized in terms of the situation calculus, in the following sense: Given a collection of actions in the STRIPS representation, you can construct a situation calculus theory defining the domain such that any plan output by the planner can be proven correct in the theory. Another example: A simulator that calculates solutions to gravitational motion by numerically solving the differential equations can be characterized in terms of a formal theory containing Euclidean space, real-valued time, and Newton's law of gravitation, in the sense that the output of the program *approximates* the conclusions of the program. (Defining this sense of "approximates" exactly is a substantial undertaking, of course.)

One major difference between compilers, STRIPS, and gravitational calculation, on the one hand, and a general commonsense reasoner, on the other, is that the former programs are doing inference in a single direction with complete information or a narrow range of partial information, whereas a general reasoner, as we have discussed, should do reasoning in many different directions using whatever partial information it has. Therefore, it is therefore more critical in a commonsense reasoner to use a widely expressive and declarative representation and a flexible inference mechanism; hence, the interest in logical representations and symbolic deduction for implementing reasoning systems. But these considerations are largely irrelevant to Hayes' argument. Note that the success of formal programming-language semantics shows that logical analysis can be valuable even when the task being studied is narrowly focussed.

The second common misconception is a little more peculiar. In fact, I can't characterize it; I can only give an example. There is a widespread impression that the only logical representations of the situation in figure 3 are something like

left-of(a,b). left-of(b,c). left-of(c,d).
red(a). white(b). red(c). blue(d).

People sometimes go so far as to assert that retrieving the fact that the leftmost object is left of the rightmost, or retrieving the fact that block E is not in this line, will take time at least linear in the number of objects.

There is, of course, not the slightest truth in this. The following are all valid logical sentences, given a suitable semantics: (Take the origin to be the lower-left hand corner of block A and the unit to be the side of that block, with axes aligned as usual.)

[3] Even in the paper "In Defense of Logic" [Hayes, 77], the argument is just that a representation should have a well-defined semantics, and that many of the "alternatives" to logic-based representations being touted at the time did not.

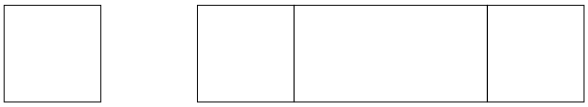


Figure 3: Blocks to be represented

place(c) = rectangle(point(3,0), point(5,0), point(5,1), point(3,1)).
red-pixel(pixel(4,0)).
empty(rectangle(point(1,0), point(2,0), point(2,1), point(1,1)).
$\forall_X$ block(X) $\Rightarrow$ $\exists_Y$ red(Y) $\wedge$ distance(X, Y) < 2.

5 Difficulties with the Manifesto

Hayes' Manifesto was much admired and widely discussed, but it was hardly followed. The committee never met, the theories were never codified. There has, of course, been a great deal of work in "qualitative physics" but this has a quite different flavor from Hayes' proposal; it is algorithmic rather than declarative, and is increasingly concerned with specialized applications rather than commonsense reasoning [Weld and de Kleer, 89], [Iwasaki, 97]. Even interpreting the manifesto fairly broadly, it would be difficult to think of more than a dozen AI researchers who have done the kind of work in physical reasoning that Hayes has in mind, while interpreting it narrowly, one could certainly argue that the manifesto and the Ontology for Liquids are the only two papers ever written that fit into Hayes' programme.[4]

No doubt the main reason for this neglect is simply that life is short, the project is large, and researchers have other things to do that seem more pressing. But, besides this, the project as Hayes outlines it has fundamental difficulties in its conception, and researchers who try to follow in Hayes' footsteps soon find themselves head to head with these obstacles. It is not really clear what, precisely, Hayes means by "naive physics". The Naive Physics Manifesto is for the most part written as if "naive physics" were a clearly defined body of knowledge — comprehensive in scope, universal across people, consistent, and essentially uninfluenced by science. More than once, Hayes claims that some specific concept or distinction is or is not a part of "naive physics", apparently in an absolute sense:

> Naive physics is pre-Galilean. I can still vividly remember the intellectual shock of being taught Newtonian "laws of motion" at the age of 11. It is interesting to read Galileo"s "Dialogue Concerning the Principal Systems of the World" (1632) where he argues very convincingly, from everyday experiences, that Newton's first law must hold. But it takes a great deal of careful argument ...
>
> I have deliberately not distinguished between mass and volume. I believe the distinction to be fairly sophisticated. [Hayes, 85]
>
> In making predictions, there is a distinction which seems crucial between events that "just happen" (such as fallings) and events which require some effort or expenditure of energy (such as rocks flying through the air). ... Such a distinction runs counter to the law of conservation of energy, and I think quite correctly so for naive physics (or we could say merely that the intuitive notion of "effort" does not exactly correspond to the physical notion of "work".)

[4] Schmolze [1986] should also be mentioned.

Now, Hayes does not, of course, actually believe in such an absolute, monolithic theory. He specifically acknowledges and discusses individual differences in the system of naive physics beliefs. Further, the first quote above at least implicitly acknowledges that an individual's beliefs may be inconsistent. (If Newton's first law can be derived from naive physics by Socratic argument and Gedanken experiments, but is also explicitly denied in naive physics, then the closure of naive physics under "reasonable argument" is inconsistent.)

The difficulties in defining an absolute "Naive Physics" include the following:

1. Naive physics is supposed to be what naive subjects believe about the physical world. But, as is well known, the concept of "belief" is ambiguous and slippery, with many different possible interpretations. "A believes ϕ" may mean that A will spontaneously assert ϕ; that A will immediately assent to ϕ; that A will assent to ϕ after Socratic interrogation; that A will assent to statements that logically entail ϕ; that the best explanations of A's actions at the knowledge level involve the assumption that A is using ϕ in the course of reasoning; or that A's actions are more sensible given that ϕ is true than given that it is false. Which is intended here?
2. The problem in (1) is made more difficult by the constraint that we are interested only in "naive" beliefs, not in beliefs that are formally taught, but that the most readily available subjects — the researchers themselves — tend to be people with substantial training in formal science and mathematics. It is not clear how we can tease out a true "naive physics" from later accretions of formal physics.
3. "Naive physics" probably varies substantially between people (though Hayes may well be right that it differs less than other branches of commonsense knowledge). Due to the vagueness in (1) and (2), it is difficult to be very detailed about this. But one can certainly see it in cross-cultural comparisons. For instance, many people in various times and places have attributed intentions and mental states to inanimate objects. In modern Western culture, this is not part of even a "naive" system of beliefs.
4. It is not clear that an individual's beliefs are consistent. It depends in part, of course, on what definition is given in (1). An inconsistent belief set cannot be expressed in a single theory in any standard logic (or indeed in most non-monotonic logics).

The result of this unclarity is that the researcher really has no way of determining whether a given concept, distinction, or rule is to be considered a legitimate element of "naive physics." Does the concept "surface area" exist in naive physics? Or the concept of an object being "awkward to handle"? Or the distinction between heat and temperature? How is one to judge? Pat Hayes [personal communication] tells a story of engaging in a two hour debate over whether a picture hanging on the wall of a room can be said to be "in" the room. Such minutiae are essentially unavoidable in this approach to formalization.

A particularly difficult issue to judge is the appropriate level of generality. Consider the rule in the cookie baking problem, "The thinner you roll the dough, the more cookies you get." Now, this fact can be expressed directly in this form. Alternatively it can be derived from the considerations that

1. The volume of the cookie dough is fixed. In particular, it is not affected by rolling it out.
2. The volume of a region is equal to its area times its average thickness.
3. The number of regions of fixed shape A that can be placed disjointly within a region R tends to increase with the area of R. (Note that this is a plausible inference, rather than a sound rule. We will consider this further in section 13.)

4. In cutting cookies out of rolled-out dough, each cookie is a cross-section of the dough on a vertical axis, and no two cookies overlap.

Or one can use rules at an intermediate level of generality (e.g. replace (2) by the more specific rule, "For a fixed quantity of malleable stuff, the thinner it is spread on a surface, the larger the area it covers"), or at a higher level of generality (e.g. derive (3) from a definition of volume as an integral.) Using the more general formulation usually has the advantages of covering more physical situations, and clarifying the relations between them, but each level of generality seems less and less "naive". How do we choose among them?

Some will argue that terms like "volume", "average" and "cross-section", which are used in our second set of rules above, are formally learned in school and therefore are not part of a naive theory. Now, certainly the more specific rule, "The thinner you roll the dough, the more cookies you get," may be one that a child learns first, before any more general formulation, and it may be a rule of thumb that someone baking cookies regularly calls upon, without doing deeper thought. But it seems to me that an intelligent person will soon see the connection between this fact and the facts that, if you want to cover a table top with books, you will do better to lay them flat and not to stack them; that a can of paint will cover a small area more thickly than a large area; and, at a further remove, that the more people are sharing a pie, the smaller each person's piece. To express the general rules that underlie these particular instances, you will almost certainly have to call on concepts that are so close to the standard ones of "volume", "average", and so on, that the distinction is hardly worth making. (Quite likely, the naive reasoner is reasoning by analogy or using case-based reasoning, rather than using an explicit generalization, but in that case these same concepts will be needed to find the dimensions of similarity between the cases. Thus the necessary expressivity of the object language is largely independent of the mode of reasoning.) Therefore, despite the association of these terms with the classroom and textbook, it seems difficult to me to justify automatically excluding these concepts, or any of the spatial concepts that we shall discuss in section 10, from a naive understanding. I should say, rather, that teaching these in the classroom is, or should be, merely a matter of putting concepts that are already understood at the commonsense level into a rigorous setting.

6 Microworlds: A modified methodology

One way out of these difficulties begins by arguing as follows: Whatever the actual content of people's individual theories, they will almost all come up with the same or similar answers over a large collection of commonsense problems. A program will achieve common sense if it gives the same answers to the same problems. Therefore, *any* theory that allows commonsense problems to be stated and solved will do. In other words, we are looking for a competence theory for solving commonsense problems. Note that we have substantially shifted our ultimate goal. Before, we were talking about expressing a body of knowledge; now we are talking about justifying a collection of inferences.

The second change that we will make is to focus on *defining a model*[5] rather than stating an axiomatic theory. The argument for this change is as follows: As discussed in section 4 our main goal in formalizing theories is to characterize or justify the actions of reasoning programs, rather than to be implemented directly as a rule base. But the relation of "justifying" a particular inference or "characterizing" a particular program is a property of a model, not of a specific axiomatization of that model. If a model can be axiomatized in two equivalent ways, the two axiomatizations support the same inferences. Therefore, our primary concern will be defining a model, and thus determining

[5]I will generally use "model" in this paper in the sense common in physical reasoning research, not in the sense of metalogic. When I need the term from metalogic, I will say so specifically.

the class of true statements and valid inferences in the model. Secondarily, we are interested in defining a formal language, which delimits expressive range, the class of facts that can be expressed. In this approach, axiomatizations are only of subsidiary interest; they help clarify the model and they are useful in verifying that a given inference is indeed supported by the model.

A third change is in the way in which the project is divided into parts. Hayes' goal is to express a theory, so a natural subset of the project is a coherent subset of the theory; that is, a cluster of concepts and axioms. The new goal is to characterize inferences, so a natural subset of the project is a *microworld*: an abstraction of a small part of physical interactions, sufficient to support some interesting collection of inferences.[6]

A few examples of microworlds:

1. The blocks world, in any of its forms.
2. The roller coaster world [de Kleer, 1977]. The world consists of a point object and a one-dimensional track in a vertical plane. The state of the world is either the position and velocity of the object along the track, or the distinguished state "FELL-OFF". The motion of the object is governed by Newton's law, with gravity and inertia.
3. Component-based electronics [de Kleer, 1985]. The world consists of resistors, capacitors, inductors, power sources, etc. connected in a circuit. The state of the world at any moment is the voltage at every node and the current through every arc. The world changes dynamically following component characteristics.
4. Rigid object kinematics. The world consisting of solid, rigid objects, constrained only by the rules that they move continuously and that two objects do not overlap.
5. Rigid object quasi-statics [Mason, 85]. The world consisting of solid, rigid objects, and externally imposed motions, constrained by the kinematic rules in (4), plus the rule that objects move only if pushed, directly or indirectly, by an external motion.
6. Rigid object dynamics. The world consisting of solid, rigid object moving under Newtonian mechanics.
7. Kinematics of solid objects and a liquid. The world consisting of solid objects and some quantity of a liquid. The motion of the solid objects is constrained by the kinematic rules in (4). The liquid is constrained to move continuously and to occupy a constant volume.
8. Kinematics of cutting solid objects [Davis, 93]. See section 7 below.

Note the difference from clusters. Of Hayes' clusters, only "liquids" is close to being a microworld, and even this would almost certainly have to be changed to "liquids and solids" (under some specified set of physical laws), as there are very few commonsense inferences that involve only liquids with no solid boundaries.

We may also contrast microworlds with reasoning architectures, such as QP [Forbus, 85] or ENVISION [de Kleer and Brown, 85]. QP and ENVISION do not incorporate any particular physical theory. Rather, each such architecture provides

a. A collection of basic ontological sorts. For instance, QP defines the sorts of time instants, time intervals, parameters, and processes.

[6] CYC, in its more recent versions [Lenat and Guha, 93] is the most notable exemplar of the use of microtheories. These are axiom-based, rather than model-based.

b. A restricted language in which certain aspects of physical theories can stated. For instance, QP supplies primitives symbols for "direct influence" and "indirect influence", which have a fixed interpretation.

c. An algorithm for carrying out certain types of inference. For instance, the QP algorithm carries out qualitative envisionment.

Thus, the development of this kind of program is orthogonal to the microworlds methodology. The microworlds approach focusses on developing specific physical theories; programs such as QP and ENVISION focus on developing techniques that apply across a range of physical theories.

Another change from Hayes' project is in the attitude toward beliefs that are commonsensical but false. These can be divided into three categories:

1. Beliefs that are approximately correct in everyday contexts. For example, the belief that a moving object will come to a halt if no force is applied. This rule, which contradicts Newton's first law, holds for most objects in most terrestrial circumstances.

2. Logical consequences of rules in (1). For example, the belief that if a torque is applied to a gyroscope, the gyroscope will rotate along the axis of the torque. This is just a special case of the general rule, "If a torque is applied to an object, then the object rotates along the axis of the torque," which holds for most objects but not gyroscopes.

3. Beliefs that are just plain wrong, without either of the above justifications. For instance the belief that an object that has been moving along a circular track will continue to move in a circle once it is free of the track [McCloskey, 1983].

A competence theory of commonsense reasoning system may well include beliefs of category (1); indeed, at some level it must, unless we plan to base it on relativistic quantum mechanics. This is justified as a trade-off of accuracy for speed and simplicity. We are therefore also likely to get beliefs of category (2), unless we can block them all using qualification conditions, which is unlikely. The question is whether there is any point in including beliefs of category (3). For Hayes' project, where the ultimate aim is a cognitive model of a naive reasoner, presumably they should be included. Likewise, if we were studying the process of *learning* physical theories, we would have to expect that sometimes the theories being considered are entirely off-base. In a competence theory of reasoning, however, since these add nothing to competence, they should be excluded. For this reason, in the new approach we speak of "commonsense physics" rather than "naive" physics.

Putting all this together, we arrive at a methodology along the following lines:[7]

1. Select a microworld: a well-defined, fairly small, range of physical behaviors.

2. Collect a corpus of inferences in the domain that are both physically correct and would be broadly agreed upon as commonsensically obvious.

3. Develop

   a. A formal model of the domain.

   b. A language of primitives with semantics defined in the model.

[7]The methodology described here is my own personal view [Davis, 1990]; however, little if any of this is original to me. Particularly significant discussions of this kind of methodology in this direction, besides [Hayes, 1978], include [McCarthy, 1968], [McCarthy and Hayes, 1969], [McDermott, 1978], [Newell, 1980], and [Charniak and McDermott, 1985]. [Halpern and Vardi, 1991] similarly argues from a shift from an axiomatic to a model-based analysis in automated reasoning.

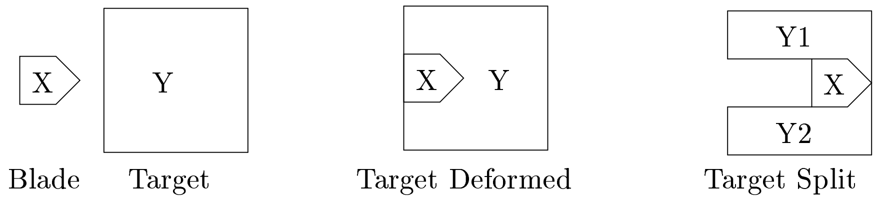


Figure 4: Mutable Object Theory

c. An axiomatization of the model expressed in the language.

4. Demonstrate that many of the inferences in (2) can be expressed in the language (3.b) and justified in the model (3.a). A formal proof from the axiomatization (3.c) may be helpful here.

5. Develop algorithms or programs that can be justified in terms of this model, and show that some significant class of commonsense inferences can be carried out efficiently.

6. Work toward broadening theories and merging multiple theories together.[8]

7 A sample microworld: the kinematics of cutting solid objects

At this point, it may be helpful to give a rather detailed description of one microworld, for illustration. The example I will use is a kinematic theory of cutting solid objects [Davis, 1993]. Relative to the state of the art in formalizing physical theories, this is a fairly complex and sophisticated example.

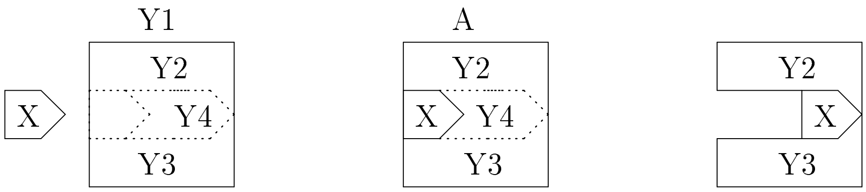
7.1 Microworld

The microworld is the kinematics of cutting rigid solid objects. That is, the world consists of solid objects moving continuously through space on arbitrary paths. The shape of any object is constant except when the objects is being cut. Objects are not created or destroyed except at the moment when one object is sliced through.

The process of cutting is modelled as if the blade annihilates the material of the target as it penetrates. When the annihilation of material leaves the target disconnected, it falls into two or more new objects (Figure 4.) This model is rich enough to support many manners of cutting: slicing through, stabbing through, filing down, or carving a cavity.

The model does not support the intuitive distinction between "cutting a small piece off of object A," where the identity of A survives in a smaller shape; and "slicing object A into objects B and

[8]John Tsotsos pointed out to me that this list should have an additional item of developing techniques to learn or acquire this knowledge. This is undoubtedly correct, but I find the idea of trying to learn this material automatically too terrifying to contemplate.



Y1 is top level.
Y2,Y3,Y4 are latent.
Y1=whole box.

A is top-level.
Y2,Y3,Y4 are latent.
Y1 is a ghost.
A=union(Y2,Y3,Y4)

Y2, Y3 are top-level
Y1, Y4 are ghosts.

Figure 5: Chunk Theory

C" where A ceases to exist and B and C come into existence. All cases where an object is split are considered in the second category, no matter how small the piece being split off.

The model does not support any theory of *dynamics,* in the sense of forces, energy, and such. For that reason, it does not incorporate any shape constraints on the blade, such as that it be sharp or serrated, or on the motion, such as that it involve sawing back and forth, as these would be arbitrary and inadequate in the absence of a dynamic theory. Similarly, the model does not incorporate the deformation of material that generally takes place in actual cutting; material is simply and irreversibly vaporized.

7.2 Ontology

We use two alternative construals of the above model. The first, more straightforward, approach construes the world in terms of objects, as above. The shape of an object O is a fluent that changes through time as O is cut and material is removed. When the shape of O become disconnected, O ceases to be "present" and becomes a "ghost"; and two new objects $O1$ and $O2$ cease to be "ghosts" and become "present." Thus, each object can undergo three types of change during its lifetime: It is originally created by being sliced off some parent object; then its shape is gradually modified as it is cut away; then it is destroyed when its shape is split.

The second construal focusses on chunks of material. A "chunk" is a physically connected piece of material; it is the part of an object that fills some connected regular[9] region. At any given moment, an object has one chunk that is "top-level", meaning that its shape is exactly the shape of the object, and many chunks that are "latent", meaning that their shape is a proper subset of the shape of the object. The latent chunks are, so to speak, waiting for a suitable cutting process to carve them out and make them top-level for their moment in the sun. A chunk of a target is "destroyed" as soon as it is penetrated by the blade. Thus the process of cutting involves the continual destruction of an infinitude of chunks which now have some of their material annihilated. At most instants, a single new chunk becomes top-level for an instance; occasionally, at the instants when the object is split, two new chunks become separately top-level. The shape of a chunk is constant. Thus, in this theory there is only one kind of change: an active chunk (i.e. one that is either top-level or latent) becomes a ghost. (Figure 5.)

[9] A regular region is one that is equal to the closure of its interior. Thus, it has no floating parts that are just two-dimensional surfaces or one-dimensional curves.

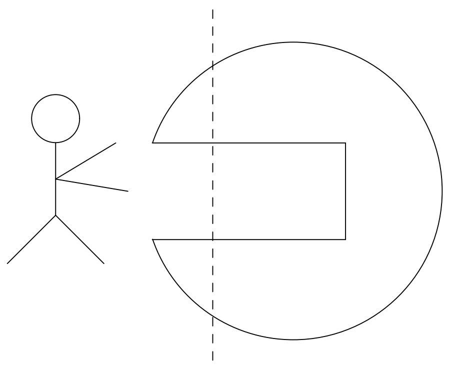


Figure 6: Carving one object or two?

The advantage of the chunk approach is that there is now only one type of change: the annihilation of material, formalized as the destruction of chunks. Sometimes this annihilation leaves a single top-level chunk, sometimes more than one, but the two essentially "look the same" from the point of view of the model. This can be useful in cases like that illustrated in figure 6. A sculptor is carving away at a pair of stone pieces, of which he can see only the nearer parts. In the object theory, this situation is difficult to describe, because he cannot know whether this is in fact one object or two; it depends on whether the two pieces are connected, which he cannot see. Worse, the two pieces may originally be a single object and then become two, when someone splits the connection behind the scene. However, assuming that the structure is fixed, it should make no difference to the sculptor whether the two pieces are connected, and in the chunk theory it doesn't. The chunks in the area visible to the sculptor are the same whether or not they are connected behind.

These two theories can be proven to be equivalent, under certain minor regularity conditions [Davis, 1993]. (These exclude scenarios in which an object is sliced infinitely many times in a finite time interval, and other such *outré* and non-physical possibilities.)

We can then define the *process* of cutting: Object A is cutting object B at time T if, for every previous time T', there was material in B at T' that A overlaps in T. (This is a minor improvement on the definition in [Davis, 1993].) Somewhat more arbitrarily, we can individuate a cutting *event*: A cutting event of B by A occurs over time interval I if A is cutting B throughout I but not throughout any proper superinterval of I.

We model time as the real line and space as three-dimensional Euclidean space. Alternative models of space and time might be possible, if they support the following concepts with suitable properties: earlier/later times, spatial regions, connectivity, rigid motions, continuous rigid motions, set difference of regions, and overlap of regions.

7.3 Language and Axiomatics

Tables 1 and 2 display languages sufficient to express the basic concepts of the two theories, and tables 3 and 4 show the basic physical axioms of the two theories. Basic geometric and temporal primitives are given their standard definitions in Euclidean space and real-valued time. The axioms are written in a sorted first-order logic. To shorten the notation, we use fluent functions as predicates with an additional situational argument. Thus, for instance, the statement "Object O is material in situation S," can be expressed equivalently either in the form "holds(S,material(O))" or "material(O, S)".

Sort	Letter
Point	X
Spatial regions (set of points)	R
Rigid mappings	M
Temporal situations	S
Fluents	F
Objects	O
Chunks	C
Either object or chunk	Q

Table 1: Logical Sorts

Temporal:

holds(S, F) — Predicate. Boolean fluent F holds in situation S.
value_in(S, F) — Function. Value of fluent F in situation S.
$S1 < S2$ — Predicate. Situation $S1$ precedes $S2$.
just_before(S, F) — Predicate. Boolean fluent F holds in an open interval ending in S.

Spatial:

$X \in R$ — Predicate. Point X is in region R.
$R1 \subset R2$ — Predicate. Region $R1$ is a proper subset of $R2$.
$R1 - R2$ — Function. The interior of the set difference of $R1$ and $R2$.
intersect($R1, R2$) — Predicate. Region $R1$ intersects $R2$.
$\emptyset$ — Constant. The empty region.
good_shape(R) — Predicate. Region R is non-empty, open, bounded, connected, and equal to the interior of its closure.
image(M, R) — Function. The image of region R under mapping M.
continuous(F, S) — Predicate. F is a continuous function of time at situation S. F is a fluent whose value in each situation is a rigid mapping.
connected_component($R1, R2$) — Region $R1$ is a connected component of $R2$.

Physical: Primitive Symbols

material(Q) — Function. The fluent of object or chunk Q being material.
placement(Q) — Function. The fluent of the mapping from the shape of Q to the place of Q.
shape(O) — Function. The point set occupied by O in a standard orientation.
cshape(C) — Function. The time-invariant shape of chunk C.

Physical: Defined Symbols

ghost(Q) — Function. The fluent of Q being a ghost.
place(Q) — Function. The fluent of the region occupied by Q in situation S.
blade_swath($S1, S2, O$) — Function. The swath cut by blades between situations $S1$ and $S2$, relative to the coordinate system attached to object O.
destroyed(S, O) — Function. Object O is destroyed at time S.
top_level(C) — Function. The fluent of chunk C being top-level.
sub_chunk($C1, C2$) — Predicate. Chunk $C1$ is (non-strictly) a sub-chunk of $C2$.

Table 2: Non-logical primitives

Definitions of Object Theory

OD.1 $\text{ghost}(O, S) \Leftrightarrow \neg\text{material}(O, S)$.
(Definition of ghost: An object is a ghost iff it is not material.)

OD.2 $\text{place}(O, S) = \text{image}(\text{placement}(O, S), \text{shape}(O, S))$.
(Definition of place: The region occupied by O in S is the image of its shape under its placement.)

OD.3 $X \in \text{blade\_swath}(S1, S2, O) \Leftrightarrow$
$\exists_{S3,OB}\ S1 \leq S3 \leq S2 \wedge OB \neq O \wedge \text{image}(\text{placement}(O, S3), X) \in \text{place}(OB, S3)$.
(Definition of blade-swath: The blade-swath between $S1$ and $S2$, relative to O, is the region swept out by all blades between $S1$ and $S2$, as measured from a coordinate system attached to O.)

OD.4 $\text{destroyed}(S, O) \Leftrightarrow [\text{just\_before}(S, \text{material}(O)) \wedge \neg\text{good\_shape}(\text{shape}(O, S))]$
(An object is destroyed at S if it existed up to S, but became disconnected or null at S.)

Axioms of Object Theory

OB.1 $[\text{material}(O1, S) \wedge \text{material}(O2, S) \wedge O1 \neq O2] \Rightarrow$
$\neg\text{intersect}(\text{place}(O1, S), \text{place}(O2, S))$.
(Two material objects do not overlap.)

OB.2 $[S1 < S2 < S3 \wedge \text{material}(O, S1) \wedge \text{material}(O, S3)\ ] \Rightarrow \text{material}(O, S2)$.
(Objects do not change from material to ghost to material.)

OB.3 $\text{material}(O, S) \Rightarrow \text{good\_shape}(\text{shape}(O, S))$.
(Material objects have good shapes.)

OB.4 $\forall_{S,O}\ \text{shape}(O, S) \neq \emptyset \Rightarrow \text{continuous}(\text{placement}(O), S)$.
(The placement of object O is continuous in any situation S where the shape of O is non-null.)

OB.5 $[S1 < S2 \wedge \text{material}(O, S1) \wedge \text{just\_before}(S2, \text{material}(O))] \Rightarrow$
$\text{shape}(O, S2) = \text{shape}(O, S1) - \text{blade\_swath}(S1, S2, O)$
(The material removed from O between $S1$ and $S2$ is the blade-swath between $S1$ and $S2$ relative to O, plus boundary points.)

OB.6 $[\text{destroyed}(S, O) \wedge \text{connected\_component}(R, \text{shape}(O, S))] \Rightarrow$
$\exists_{OR}\ \text{shape}(OR, S) = R \wedge \text{placement}(O2, S) = \text{placement}(O1, S) \wedge$
$\text{just\_before}(S, \text{ghost}(OR)) \wedge \text{material}(OR, S)$.
(If O becomes disconnected or null at S, then each of its connected components become material.)

OB.7 $[\text{material}(O, S1) \wedge \text{ghost}(O, S2) \wedge S1 < S2] \Rightarrow \exists_{S3 \in (S1, S2]}\ \text{destroyed}(S3, O)$
(An object turns from material to ghost only if it is destroyed in the sense of OD.4.)

OB.8 $[\text{ghost}(O, S1) \wedge \text{material}(O, S2) \wedge S1 < S2] \Rightarrow$
$\exists_{S3,O3}\ \text{destroyed}(S3, O3) \wedge S1 < S3 \leq S2 \wedge$
$\text{connected\_component}(\text{place}(O, S3), \text{place}(O3, S3))$.
(An object can come into existence between $S1$ and $S2$ only if it is a connected component of some object $O3$ that is destroyed at some $S3 \in (S1, S2]$.)

Table 3: The "mutable objects" theory.

Definitions in Chunk Theory

CD.1 ghost$(C,S) \Leftrightarrow \neg$material$(C,S)$.
(Definition of ghost: A chunk is a ghost iff it is not material.)

CD.2 place$(C,S) =$ image(placement(C,S), cshape(C))
(Definition of place: The region occupied by C in S is the image of its shape under its placement.)

CD.3 sub_chunk$(C1,C2) \Leftrightarrow$
$\exists_S$ material($C2$,S) $\wedge$ place$(C1,S) \subseteq$ place$(C2,S)$.
(Definition of sub-chunk: $C1$ is a sub-chunk of $C2$ iff $C1$ occupies a subset of $C2$ in some situation where $C2$ is material.)

CD.4 top_level$(C,S) \Leftrightarrow$
[material$(C,S) \wedge \forall_{C1}$ [material$(C1,S) \wedge$ sub_chunk$(C,C1)] \Rightarrow C1 = C$].
(A top-level chunk is a maximal material chunk relative to the sub-chunk relation.)

Axioms of Chunk Theory

CH.1 good_shape(cshape(C)).
(Chunks have a good shape.)

CH.2 [good_shape$(R1) \wedge R1 \subseteq$cshape$(C2)] \Rightarrow$
$\exists^1_{C1}$ $R1$=cshape$(C1) \wedge$ sub_chunk$(C1,C2)$.
(Every reasonably-shaped subregion of a chunk is a chunk.)

CH.3 continuous(placement(C), S).
(The placement of chunk C is continuous in every situation.)

CH.4 [sub_chunk$(C1,C2) \wedge$ material$(C2,S)] \Rightarrow$ material$(C1,S)$.
(A sub-chunk of a material chunk is itself material.)

CH.5 [sub_chunk$(C1,C2) \wedge$ material$(C2,S)] \Rightarrow$
placement$(C1,S) =$ placement$(C2,S)$.
(A sub-chunk of a material chunk has the same placement.)

CH.6 material$(C,S) \Rightarrow \exists_{C1}$ top_level$(C1,S) \wedge$ sub_chunk$(C,C1)$.
(Every material chunk is a sub-chunk of a top-level chunk (possibly itself).)

CH.7. [material$(C1,S1) \wedge$ ghost$(C1,S2)] \Rightarrow$
[$S1 < S2 \wedge$
$\exists_{S3,C2}$ $S1 < S3 \leq S2 \wedge \neg$sub_chunk$(C1,C2) \wedge$ top_level$(C2,S3) \wedge$
intersect(place$(C1,S3)$, place$(C2,S3)$)].
(A material chunk $C1$ can only turn into a ghost if its interior is penetrated by a top-level chunk.)

CH.8 [top_level$(C1,S) \wedge$ top_level$(C2,S) \wedge C1 \neq C2] \Rightarrow$
$\neg$intersect(place$(C1,S)$, place$(C2,S)$).
(Two top-level chunks cannot intersect.)

Table 4: Chunk Theory

7.4 Inferences

The model supports exact predictions: given the positions and shapes of all the objects at the start of a time interval, and given the motions of all the objects throughout the interval, predict the identity and shapes of the final objects at the end of the interval. This is the kind of prediction that is carried out in CAM machining programs [Ji and Marefat, 97].

It also supports kinematic inferences of other kinds. For example, [Davis, 93] gives the proofs of the following statements:

- A blade that starts outside the target cannot carve a purely internal cavity inside the the target.
- If the blade is restricted to linear motions, then carving out a k-face convex polyhedron requires at least k separate cutting operations.

In our original scenario 2, of the cookie dough, this model supports most of the inferences one would want to make about cutting the dough with cookie cutters, assuming that the dough is otherwise rigid during the cutting process. For instance, one can conclude that, if the dough is cut in the center by a cutter that is a simple, non-closed, curve, then no cookie has been separated out. One can conclude that, if the horizontal projections of two cuts with ordinary cutters overlap, then the cookies cut out are the connected components of the intersection and the set differences of the two regions within the cutters.

7.5 Observations

The strongest aspects of this formalization are, first, its generality, the fact that slicing, stabbing, and filing can all be treated together; and, second, its clarity; potential confusions are almost entirely resolved. If you try just to write down everything you know about "cutting", you are apt to find that there are a large number of issues to resolve, and that it is difficult to ensure that you are resolving them all consistently. This approach takes care of all these.

Moreover, these models seem cognitively plausible as far as they go. It seem very natural to think about individuated objects being gradually shaved away by a cutting process; it seems almost as natural to think about chunks of material, particularly when the extent of the object is either unknown, as in figure 6, or is very much larger than the region being operated on. The theories are certainly rather abstract and bloodless, mostly, I suspect, due to the absence of any dynamic theory. A lot of one's experience of cutting has to do with the forces and motions involved in sawing, stabbing, and so on, and these have all been abstracted away in this microworld.

8 Advantages of microworlds

In this section and the next, we discuss the strengths and flaws of this new approach of constructing microworlds to formulate a competence theory. Regrettably, the distinction between strengths and flaws is not as clear-cut as one might like. Some apparent strengths may actually be flaws; some apparent flaws may actually be just hard problems that would be encountered in any methodology.

The first and foremost advantage of the competence theory approach is that it takes us away from the painfully vague problem, "Is concept / distinction / fact X part of naive physics?" and replaces it by the much more hard-edged, pragmatic, bottom-line, engineering-type question, "Is X useful over a given class of inference?" For instance,

- Is an elastic collision between solid objects an instantaneous event, involving an instant change in velocity, or is it a prolonged process, involving an extended period of contact, a continuous change in velocity, and a deformation of the objects involved?
- Can a physical object be truly a point, or a curve, or a surface?

It is difficult to justify or even to assign meaning to a claim that one or the other of these is the "true" naive view. It is much easier to say that one or the other is an adequate model over a given class of inferences. Discussions such as that mentioned above about whether a painting on the wall is in the room can be avoided. What is actually going on, geometrically and physically, is quite clear enough and easily described. How you choose to define "in", whether you want to define the spatial extent of the "room" to include the walls, and whether you want to define the "walls" to include the painting (is the painting *part of* the wall or merely attached to it?) are comparatively unimportant and arbitrary decisions about the symbols "in", "room" and "wall".

This freedom from worrying about whether concepts are truly naive comes about primarily because, while Hayes' project requires that naive conclusions be drawn from naive premises, our project require only that naive conclusions be derivable; the premises need not be formulated in naive terms. Therefore, whereas Hayes' project requires that every concept be examined for its true naivity, and rejected if it is not genuinely naive, for us it suffices to have a large collection of naive conclusions. To carry out our project, in other words, it suffices to be able to generate a large collection of inferences that are unquestionably commonsensical; we never have to decide of a given inference that it is *not* commonsensical.

The problem of finding an appropriate level of generality, which we considered above, is likewise considerably clarified in the new approach. To attain maximal inferential power, one always goes to the highest level of generality that has any justification within the scope of the microworld. For example, in the cookie-cutter example, one can derive the rule from a very general theory of volume of regions together with the physical rule that the volume of the dough remains nearly constant while being rolled out. This general theory of volume will serve for many other inferences that involve reshaping of malleable, incompressible material, so it is advantageous to do this at a general level. On the other hand, there is probably nothing to be gained from abstracting further to the general notion of a Lebesgue integral in a general measure space; within commonsense physics, there will be no interesting generalizations to be obtained from this more abstract notion.

Once we are using microworlds in a competence theory, it becomes almost irresistibly tempting to consider competence over particularly interesting limited classes of inferences as a final goals in themselves. One can therefore contemplate the possibility of using multiple, mutually inconsistent, microworlds for the same phenomena, depending on the scope of inferences being considered and the precision required. For instance, in section 6 we enumerated three different theories of immutable solid objects: pure kinematics, quasi-statics, and Newtonian dynamics. Each of these (as well as other theories, such as the theory of deformable solid objects) is a useful theory under different circumstances. This is more difficult to justify in the project of expressing "naive physics", where we are presumably looking for a coherent universal theory.

This ability to consider microworlds for limited purposes has a number of advantages. First, it makes the analysis much easier; we can focus in on getting some particular class of inferences to work without worrying how these will fit with all the rest of naive physics. Second, it allows much closer ties to practical applications. Most practical AI physical reasoning programs work within a quite limited scope. For instance, many of the programs that do mechanical reasoning [Joskowicz and Sacks, 91], [Faltings, 87] work within the microworld of solid object kinematics or some small extension of it. As we shall argue further below, this tie to practical applications is very valuable for a number of reasons. Third, as the work on automated modelling [Nayak, 94] has shown, there can be considerable computational advantage to being able to choose, for a given problem, models of the correct level of precision and detail, so that correct answers can be reached without unnecessary excess computation. The study of alternative microworlds connects directly to this kind of study. (In section 14 we will consider some limitations of this point of view.)

Focussing on the model rather than the axiomatization has the usual advantages of making it much easier to ensure consistency and to avoid unintended consequences. As discussed in section'—refsec-microworlds, a concrete extensional model is necessary consistent and precisely defined, and so avoids a lot of the conceptual inconsistency and incoherence that can arise in the axiomatic approach.

9 Dangers and difficulties of microworlds

This revised approach does not, however, take us out of the woods and cure all of our methodological difficulties. On the contrary, though some difficulties are alleviated from Hayes' original formulation, many are no lighter, and some are worse.

The chief problems are these:

- Commonsense reasoning is not an autonomous task domain.
- It is hard to find natural sources for commonsense inferences in a single microworld.
- The number of potential microworlds is vast, and the methodology provides no guidance for choosing between them.
- The focus on microworlds rather than axioms encourages (a) excessive specificity; (b) overemphasis on mathematical abstraction and elegance; (c) overemphasis on deductive reasoning.
- There is no easy way to integrate microworlds.
- The method involves endless hairsplitting of essentially vacuous issues.

We will elaborate on each of these individually.

9.1 Not a task domain

The central objective in the new approach is to develop a competence theory for commonsense physical reasoning. But a competence theory must describe competence in some particular task, and "commonsense reasoning" is not, in itself, a task.[10] That is to say, it is not a cognitive activity that takes place by itself in people, or that would be of any value taking place by itself in a computer; it is an aspect of other cognitive tasks, such as planning actions, natural language understanding, expert systems, and so on. Moreover, the connection to commonsense reasoning is the most poorly understood aspect of these tasks, and at the current stage of understanding, such systems are very rarely improved by any attempt to incorporate commonsense reasoning.

Commonsense inference is thus an ill-understood module of a much larger task. It is therefore very difficult to be sure what the inputs and outputs of this module should be; that is, to decide how a commonsense inference should be formulated in order to serve the purposes of these larger tasks. In considering commonsense inference for a natural language processor, for example, it is difficult to

[10] It is noteworthy that in the paradigmatic case of a competence theory, natural language syntax, the Chomskian linguists have felt obliged to focus on a very narrow and artificial task, that of judging grammaticality, rather than think about more ecologically valid tasks, such as producing or comprehending natural language. It might be worth considering whether some analogous task could be found in our domain.

know which aspects of the inference are part of the purely linguistic component and which parts are part of the commonsense reasoner. It is also difficult to know what is involved in "understanding" a given text.

For instance, consider the text, "Use a rolling pin to roll out the cookie dough on a flat surface that has been covered with flour. Then cut it into pieces with a cookie cutter." Interpreting this text involves making the inference that "it" refers to the dough rather than the rolling pin, the surface, or the flour. This inference requires a combination of linguistic rules and commonsense reasoning. But it is not easy to tell what commonsense inference, precisely, is involved here. Do we want to infer that it is difficult to cut a rolling pin, or that it is unusual to do so, or that doing so will serve no purpose in the recipe? In the same way, it is difficult to know what is needed to achieve "understanding" of the text. Does the task of natural language understanding, as such, require inferring that the surface is horizontal? or that the cutter is moved downward through the dough to the surface? (Translation into another language often does require such knowledge, in order to choose the proper spatial terms.) In short, the problems of what the representation of a text should be and of how world knowledge can be used in linguistic analysis are very obscure, and therefore it is difficult to get guidance for commonsense reasoning and representation from linguistic examples.

Similar ambiguities appear in relating commonsense reasoning to robotics. Here they take the form of the uncertainty of knowing what a high-level plan looks like, and how it relates to low-level robot programming. Suppose we want to build a robotic system that can carry out the cookie dough plan. Then the system effectively infers the statement "If program P is carried out on robot R in situation S, then the goal of having cookies will probably be achieved." This is not, in itself, a statement analyzable within a commonsense physical theory, as P mostly consists of a lot of low-level robot-specific instructions governing manipulation, vision, and hand-eye coordination. It is not at all clear what higher-level plan should be the subject of commonsense reasoning, or what statements should be inferred about such a plan. The general issue here is the problem of determining what issues should be addressed in high-level planning and what is the form of a high-level plan. Again, these difficulties make it hard to use robotic programming as a guide to commonsense reasoning.

Let me clarify the problem here by contrasting commonsense reasoning with two other hard tasks. Automatic dictation, from voice to manuscript, is hard, but at least we know the form of the input (an acoustic string) and the output (a sequence of characters) and we have an unlimited collection of examples where we know that a correctly working program will produce output O for input I. Fluid flow analysis for rocket testing is a hard module to build, but again we know that the input should be the boundary conditions for the relevant PDE and a specification of the desired precision, and that the output needed is a field of fluid flow of that precision. The difficulty with commonsense reasoning is that there are very few instances where we can be really sure what the input and the output should be.

9.2 No natural sources for single microworlds

Once we have chosen a microworld, we have to find a collection of inferences within that microworld as a testbed. It is important that the collection should well represent the range of commonsense inferences in the domain, in terms of the physical phenomena considered, the types of partial knowledge, and the directions of inference. If the collection of inferences is too narrow, then it is likely that the model developed from them will be too weak or the language too inexpressive.

The problem then is, how does one assemble a suitably broad collection of commonsensical inferences within a given microworld? The best way would be to choose a task that is easily carried out by naive subjects, such as vision or language interpretation, and collect the commonsensical inferences within this microworld involved in that task.[11] But this is hard to do, as discussed in the previous section. Reasoning for expert systems, or processing of specialized natural language text, or planning for special-purpose robotics often stays largely within a small microworld, but rarely covers the range of commonsense inference; these tend to be confined to a few types of inference (e.g. prediction) and to very few types of partial knowledge. Within these confines, they go far beyond commonsense reasoning in specialized techniques and knowledge (otherwise, they wouldn't be *expert* systems.) Natural language processing of general text and planning in rich environments uses many more types of inference, but only occasionally do these fall within the chosen microworld.

The method of exploring variants, advanced in section 1, often yields a collection of interesting problems, but it has a number of built-in biases: it tends to favor prediction problems over other directions of inference; it tends to favor fairly complete specifications; and, being generated by the researcher herself or sympathetic colleagues, it can easily be biased toward conforming to the theory the researcher has in mind. Also, a researcher who has thought for a long time about a given microworld may well tend to exaggerate how easily naive subjects can make certain inferences, so she may include as commonsensical inferences that are in fact quite difficult.

For example, suppose we want to evaluate how well the model and language of section 7 characterize commonsense inference about cutting solid objects. How can we go about such an evaluation?

A claim of adequacy must be that a significant fraction of the commonsense inferences in this domain can be justified in this theory. How do we find or define a space of typical commonsense inferences within this microworld? We can look at the inferences that a CAM machining program is implicitly carrying out, or the additional inferences that it would be useful for such a program to carry out. But most of these are of the form "To create hole H in object O, move cutter C through path P", which can all be satisfied by a substantially simpler model of cutting, such as one in which each operation with the cutter is taken to be atomic; and by a simple language, such as one in which all geometric descriptions are exact. Most of the other inferences fall outside the microworld, such as restrictions of the thinness of the parts that can cut out of a given material with a given cutter. We can look at natural language processing of a technical text describing machining. This will probably yield a slightly broader class of inferences within the microworld than the CAM program, but still a quite restricted one. We can look at unrestricted text, but how frequently does any interesting issue in cutting solid objects arise in novels or in the newspaper?

So the question, which is naturally often raised, of how this theory could be implemented, is one that I can hardly answer, because I have no idea what such an implementation would be supposed to *do*. I could implement a predictive program that takes exact initial shape descriptions and description of motions and output final shape descriptions, but this has been done by the CAM people much better than I could do it. I could set a general-purpose complete theorem prover on the axiom set of tables 3 and 4, but for a theory of this complexity, I would not expect an answer in reasonable time to any but the most trivial queries. What I am looking for is a inference engine that will work efficiently over the space of commonsensically obvious inferences, but I don't know what that space *is*, let alone how to design an inference engine for it.

I am not, of course, arguing that commonsense inference has no practical application. I am arguing that the practical applications are apt to be few until we have gotten far past simple microworlds to very broad theories. A program that could do general commonsense reasoning would be of immense value; a program that could do physical commonsense reasoning, broadly interpreted, would be of very great value; but a program that can do commonsense reasoning about cutting solid objects, or similarly narrow domains, would be of very little value. Therefore, it is very difficult to know what any of these programs should do about cutting solid objects. We don't know what a

[11] Interestingly, the original plan for CYC [Lenat et al., 86] was to express the background knowledge needed to understand encyclopedia articles (hence the name); they later report that "that use of external written materials has become increasingly rare." [Lenat and Guha, 93].

program that only does commonsense reasoning about cutting solid objects should do, because there is almost nothing useful that it can do. We don't know what kinds of reasoning about cutting solid objects a general commonsense reasoning program would be called on to do because it will only very occasionally called upon to carry out an inference that is both non-trivial and lies entirely within this microworld.

This problem is serious, not just because the absence of short-term payback makes it difficult to attract the interest of colleagues, students, and funding agents, though these considerations are not to be sneezed at. Far more importantly, it means that there is almost no way to guide research in microworlds to evaluate what progress is being made, except for the judgment and taste of the researcher [McDermott, 1987]. We have to work almost blind until the work is almost complete.

9.3 Innumerable microworlds

Hayes' project is large but, at least in principle, it is finite; once the knowledge of all naive physics has been formalized, the project is done. Our project, by contrast, is infinitely open-ended or nearly so; one can continue to make up and analyze new microworlds forever, by slightly varying the set of assumptions involved. For example, there is an endless collection of variations on the blocks world: Blocks may stack in towers one on one, or they may be rectangular of varying sizes, or they may have more general shapes; time and space may be continuous or discrete; there may be one hand or many hands, and, if many, they may work one at a time or concurrently and they may interact in any of several ways; and so on. This is useful for the teacher giving a class in KR who needs simple examples to assign, but for the researcher, these are mostly useless distractions. The methodology described above does not give one any clue as to when the analysis of a new microworld is worthwhile. The choice of where to invest energy is left entirely up to the judgment of the researcher, and KR research has always been remarkably apt to leave the great ocean of truth undiscovered, while crowding around an empty Chlorox bottle on the beach.

9.4 Excessive specificity

A microworld is, so to speak, an entire alternative universe that approximates or abstracts the real one. In formulating a microworld, therefore, it is often difficult to avoid ontological overcommitment; being overly specific merely for the sake of having a well-defined model.

Suppose, for example, we want to describe cell division. At the beginning of the process, there is one cell; at the end, there are two cells. For simplicity, it is certainly easiest to say that up to a certain time there is one cell, called A, and after that time, there are two cells, called B and C. Clearly, however, isolating the moment in the process that divides one from two is entirely arbitrary and pretty pointless.[12] So we would like to be agnostic about this. In the axiomatic approach, this agnosticism is very easily attained, through the following axioms:

1. At the beginning of cell division, A exists.
2. At the end of cell division, B and C exist
3. At all times during cell division, exactly one of two possibilities holds: (1) A exists; (2) B and C exist.
4. It cannot be that B and C exist at one time and then A exists at a later time.,

[12]Deciding which state holds at that exact moment is doubly arbitrary and pointless.

By contrast, the whole spirit of the microworlds approach militates against this kind of agnosticism. Characteristically a microworlds approach will feel obliged to define a criterion for the temporal individuation of cells, and this criterion will impose a unique solution to the question of the dividing point. Without such a criterion, the only way to achieve conditions (3) and (4) is to define the temporal lifetime of one cell in terms of the lifetime of another, and this kind of recursion loses many of the advantages of microworlds, such as the easy guarantee of consistency. I'm not saying it can't be done; I'm saying that someone working within the microworlds methodology is much less likely to adopt such a solution or to be satisfied with it. Another, more complex, example of this same tendency will be discussed in section 9.7.

The model-based methodology also pushes toward excessive specificity and concreteness in the concepts considered. The focus is on concepts that are easily characterized in terms of their spatial/temporal/material aspects to the exclusion of more nebulous but important concepts attached to causality and teleology. Consider the following inference

> If you cut through an object anywhere near the center, you will probably destroy its functionality.

The inference is important, true, and commonsensically obvious, but is likely to be omitted in a model-based theory, because of the difficulty of defining "functionality". It is also unlikely to be found as a sample commonsense inference by the method of proposing variants, because it is too general.

9.5 Excessive mathematization

Similarly, the model-based methodology leads to an excessive interest in constructing elegant and minimal mathematical models rather than expressive, messy models. For example, the kinematic theory of cutting solid objects presented in section 7 is elegant and simple, easily stated and formalized, covering a wide range of phenomena with a few rules.

The dynamic theory of cutting solid objects, by contrast, is complex, haphazard, and incomplete. Consider the range of motions, forces, and behaviors involved in slicing through butter, sawing wood, driving a nail, screwing a corkscrew, and drilling a hole. A model that characterizes all these fully at the commonsense level will necessarily involve a large number of separate rules and constraints governing these separate common cases. (The theory at the atomic level is simple, but there the structural representations needed to describe these various scenarios is very complicated.) Moreover, these rules and constraints are not disconnected arbitrary facts, but are deeply interconnected. For instance, anyone who has observed the processes of butter being sliced and of wood being sawed will expect, from the nature of the processes and the materials, that butter can be sliced more thinly that wood can be sawed. But it is not easy to find the general rules that give rise to that expectation.

The researcher who wants to move forward producing models will therefore tend to avoid this kind of microworld, as these models are, in every respect, harder to develop. The ontology and language are much richer; the theory is much more complex; it is hard to be sure that the various constraints and rules are mutually consistent; it is hard to be sure that all cases have been covered. Paradoxically, one suspects that this kind of model are also harder to "sell" as legitimate research; they look like a mere translation of random obvious statements into formalese. In fact, the immense gap between a mere translation of random statements and a coherent theory is no less in a complex theory than in a simple one, but the coherence of the complex theory is harder to achieve, to convey, and to grasp.

In fact, as the microworlds become more complex, the need for complex systems of constraints on the models means that the distinction between the axiomatic approach and the model-based approach tends to vanish. Each of these constraints is, effectively, an axiom; the difficulties of dealing with the constraints are almost the same as the difficulties of dealing with a set of axioms; and the advantages of a model-based approach over an axiomatic approach, in terms of clarity and of easily-verified consistency are much diminished.

Having constructed elegant models for simple domains, the next temptation is to spend time proving neat theorems about them, or in them. These are often of doubtful relevance. A twenty-two page proof that two theories of cutting are mathematically equivalent [Davis, 93] certainly does not represent any cognitive activity that anyone (except myself) has ever carried out, nor any computational activity that any program is ever likely to carry out. Now, of course, I can and do justify such research in terms of the methodology itself: a program that can reason flexibly about cutting must be based on a good model of cutting; the two models potentially have different advantages as regards automated inference; if we want to use them both, we should understand the relation between them; hence, it is of value to know that they are in fact equivalent. Which is all very well, but all the same the gap between application and research has gotten rather large.

This mathematizing tendency also affects the formulation of queries. In section 5 , we suggested that the special rule "The thinner you roll the cookies dough, the more cookies you can cut out," could be deduced as a consequence of more general geometric rules, plus rules that the cookie dough has fixed volume, and that cutting out cookies corresponds to dividing the region of the dough into vertical cylinders with some fixed cross-section. But this "generalization" fails to capture the causal direction of the general rule, the fact that the baker can choose how thick to roll the dough and where to cut the cookies, and that these choices determine the number of cookies obtained. By contrast the geometric rules are atemporal; they would equally apply to a case where someone was assembling a mass of cookie dough out of cookie pieces, and where the choice of the number of cookie pieces would determine the eventual volume of cookie dough. A large part of mathematical training involves making this kind of abstraction automatic; it eventually becomes so much second nature that perceiving the distinction between the original rule and its abstraction requires a conscious effort.

9.6 Too much stress on deduction

Being centered around semantic consequence, the microworlds approach tends to focus exclusively on deductive reasoning. It can, perhaps, be extended to types of plausible reasoning based on a strong semantic model, such as circumscription or probabilistic reasoning (section 13) but would be very difficult to integrate with such theories as default reasoning, reasoning by analogy, case-based reasoning, and so on.

9.7 Combining microworlds

One advantage to using clusters of axioms is that, if you get the axioms for each cluster "right" — and I put this in scare quotes because it's not clear what "right" means in this context, or even if it has any coherent meaning — but supposing that there is some coherent meaning, if you get the axioms for each cluster right, then clusters can be combined just by taking the union of the two clusters, and perhaps adding some additional connecting axioms. Certainly, two "right" clusters of axioms are by definition mutually consistent. By contrast, the notion of "combining" two microworlds or "extending" a microworld is not usually well defined. It may be that some of the ideas or the axioms transfer from one microworld to a richer microworld, but only occasionally will one microworld be part of another in any formal sense, and even then, what the formal relation between a more restricted theory and a broader theory is varies from one case to another.

Let us first consider two examples of theory extensions that can be characterized formally. The kinematic theory of rigid solid objects can be extended to a dynamic theory by adding mass, force, momentum, and so on, and imposing Newton's laws. This is what Giuchiglia and Walsh [1992] call a "theorem increasing" extension; the language is richer and the axioms of dynamics are a strict superset of those of kinematics. It is also, correspondingly, "model[13] decreasing" [Nayak and Levy, 94]; if H is a history consistent with the dynamic theory then the "projection" of H obtained by eliminating all aspects of the history except shape and position is consistent with the kinematic theory.[14]

The kinematic theory of rigid solid objects can also be extended to the kinematic theory of cutting rigid solid objects, described in section 7. This is a model-increasing extension; any history consistent with the non-cutting theory is also consistent with the cutting theory. Correspondingly, it is a theorem-decreasing extension. This seems a little odd, as the cutting theory contains all kinds of axioms and inferences about cutting that don't apply in the non-cutting case, but actually these are all vacuously true in the non-cutting case. For instance, it is true in a non-cutting theory that if a knife cuts through an apple, the apple will be split into two parts, because the antecedent of the implication is necessarily false. (Note that statements like "You can cut through an apple with a knife," are not consequences of the theory in section 7.)

But both these cases are atypical, in that there is a formal relation between the two theories. What is much more common is theories that are incommensurable and uncombinable. Let me discuss an example that has been fretting me for some years. I have a theory of cutting rigid solid objects. I also have a theory of strings, presented briefly in [Davis, 95b]. The form of this theory is determined by the following considerations:

A. The length of a string is constant.

B. Strings are very flexible.

C. It is tempting to make strings one-dimensional curves, but that creates difficulties. For instance, if two strings touch one another, or one part of a string touches another part, then, if the strings are truly one-dimensional, it becomes very difficult to specify which string is on which side. Consequently, it becomes difficult to fix the rules so that one string cannot pass through the other. It is much easier to specify a reasonable physics if strings are required to be fully three-dimensional objects, though thin.

D. The cross-section of a string is pretty much constant.

E. We wish to abstract away the details of the composition of the string, which varies from one string to the next, and focus on the external characteristics, which are very much the same from one string to another.

To accommodate these constraints, I proposed the following kinematic theory of strings and solid objects:

> A string is characterized by its length L and its radius R. At any given moment, the *core* of the string lies on a curve C of arc-length L. The extension of the string lies on all points of the form $\mathbf{q} + \Delta\hat{N}$ where $\mathbf{q}$ is a point in the core C; of f; $\hat{N}$ is normal to the curve C at $\mathbf{q}$; and $\Delta \leq R$. The string observes the following constraints:

[13] This is "model" in the strict metalogical sense.

[14] If you allow the imposition of arbitrary external forces and impulses as boundary conditions then a version of converse also holds: Given any (piecewise twice-differentiable) motion satisfying the kinematic constraints, there is some way of imposing external forces so that the objects execute the motion in the dynamic theory. At this point, the question of which, if either, direction is "theorem decreasing" and "model increasing" becomes rather murky.

- The string moves continuously.
- The string does not overlap any solid rigid object.
- The string does not overlap any other string.
- The string does not overlap itself. That is, there cannot be two distinct points $\mathbf{q}_1$ and $\mathbf{q}_2$ on curve C; two normals $\hat{N}_1$ and $\hat{N}_2$ to C at $\mathbf{q}_1$ and $\mathbf{q}_2$; and two quantities Δ_1, $\Delta_2 < R$, such that $\mathbf{q}_1 + \Delta_1 \hat{N}_1 = \mathbf{q}_2 + \Delta_2 \hat{N}_2$

This theory is reasonably straightforward, and integrates directly with the kinematic theory of rigid objects. It supports inferences such as, "If string A is looped, with one end flush against the other, and string B is likewise looped, and the two cores are topologically linked, then the two strings cannot be separated from one another while keeping them both looped." The topological part of this proof is not easy, but the physics is simple.

The problem now is, how can the theory of strings be combined with the theory of cutting? The difficulty is that halfway through the process of cutting the string, the string has a notch that has been vaporized out of it. The theory of strings, as stated above, applies only to strings with circular cross-sections everywhere, and it is not at all easy to extend it to handle a string with a notch removed.

Now there may be a good, or at least a deep, reason for this difficulty. The model of string as a uniform tube is an abstraction of many different string-like substances: woven string, braids, single fibers, metal wires, rubber-coated wires, even linked chains. The abstraction is reasonable across a wide range of behaviors, but it falls apart in scenarios that probe the internal structure of the string. (By definition, of course: a scenario that distinguishes one internal structure from another is precisely one in which the internal structure cannot be abstracted away.) Chief among these is cutting or partially cutting the string; what happens when you cut halfway through a string is quite variable, depending on what the string is made of. Hence, it is not surprising that modelling cutting string is not a simple extension of modelling string.

On the other hand, cutting string is not, after all, a very esoteric activity, and the fact that, when you cut a string, you end up with two shorter strings is one of the best-known and most important properties of string. Two related reactions to the above difficulty come to mind immediately. The first is that we don't care what's going on in the middle of cutting string; all we care about is the end result. The second is that we don't generally care about strings that have been halfway cut through; when we start to cut a string, we usually complete the job.

The first of these reactions is actually a fallacy, based on the ease with which human reasoners solve and therefore ignore the frame problem. After all, cutting string does not create a physics-free zone, and we would care very much if string, while it was being cut, spat forth a poison that was fatal on contact. So the reaction "we don't care" is presupposing some very strong constraints on the behavior of the string while being cut that carry over from before it was cut, and our problem is precisely to state these constraints in a way that integrates with the rest of our theory.

The second reaction is more productive. We could look for a model in which the string is never partially divided, by positing that the string splits in two as soon as it is penetrated by the blade. This can be accommodated in chunk theory by observing that, unlike soap or marble where any reasonable subset can be carved out, strings can really only be cut straight through. (If you do manage to cut a string lengthwise, then what you get may very well not be a string.) Therefore, if we take a "chunk" to be "something that can potentially be cut out of the material," then the chunks in the string are precisely lengthwise segments of the string. If we apply our rule from chunk theory that a chunk vanishes as soon as it is penetrated, then what we get is precisely the above model, that the string is split as soon as the blade enters it. (Chunk theory also allows a more elegant expression of the last rule above, that the string does not overlap itself.)

This theory seems elegant enough, and it does the right thing for almost all cases of cutting string, so in that sense it is a reasonable competence theory.[15] Unlike the microworlds we have looked at before, however, the description here is never either true or plausible; strings do not split in two the instant that the knife enters them, and one does not imagine that they do. Moreover, on the rare occasions when it is obvious that the knife will partially cut the string but not wholly, this gives a prediction that is neither right nor plausible.

What we have done, in short, is to construct a concrete model of the process of cutting, which has the correct starting and ending behavior for completed cuts and the correct interaction during cutting with the rest of the world (i.e. none). Then this model will do the right thing as long as we never have to reason about incomplete cuts or about the state of the string during cutting. The fact that it is easier to construct such a overly specified model rather than just characterize correctly the starting and ending states and the interaction with the rest of the world is a fine example of how the model-based methodology pressures one into overly specific models (Section 9.4.)

9.8 Hairsplitting

By this stage of the paper, few readers will need of more illustrations of this point! Back in section 8, we were patting ourselves on the back that we could avoid two-hour discussions on the meaning of "in", but though that particular vacuous argument is avoided, many others come in to take its place. The kind of precision needed in this kind of analysis seems to require inescapably that all kinds of borderline cases and anomalies be resolved.

In the case of real borderline cases — Is a platypus a mammal? Is glass a solid? What is an impulse? — this is somewhat tolerable, as scientists and engineers who study this kind of issue also spend serious work doing this kind of resolving of borderline cases. Even here, one's intuition is that human commonsense reasoning is distinguished by its willingness to admit the existence of borderline cases, and its non-insistence on tying all these down; and one would like the theory of automated commonsense reasoning to be similarly flexible. What is truly intolerable, however, is the amount of time and effort that must be spent in resolving purely hypothetical and imaginary borderline cases and anomalies, just for the sake of having clear-cut definitions and models — When you turn on a light, is it on or off at the exact dividing moment? Do objects occupy open or closed regions in space? What happens if an object is sliced simultaneously by infinitely many blades? No scientist or engineer would dream of wasting her time in this way; here we are in company only with mathematicians and philosophers. Mathematicians have it comparatively easy; the hairs only have to be split when choosing definitions, not when proving theorems; mathematics tends to have few definitions and many theorems; and hairs can be split along any lines that seem most convenient. By contrast, we spend much more of our time defining concepts and models, and we are under pressure to make our definitions more or less fit with commonsense concepts. Philosophers have it even worse than we do; rather than analyzing straightforward concepts like cutting string, they are trying to deal with Truth, Justice, and Beauty. On the other hand, of course, the reward for their efforts is a better understanding of Truth, Justice, and Beauty, whereas the best we can hope for is a better understanding of how to formalize cutting string.

[15] I have not worked through this theory carefully, and so there may be some technical problems that arise. It is a little worrisome, for instance, that in this theory a solid object exists over a time interval that is closed on the left and open on the right, while a string exists over an interval that is open on the left and closed on the right. My guess, though, is that this does not raise any real difficulties.

10 Spatial Reasoning

The remainder of this paper discusses a number of issues in commonsense physical reasoning that I believe are critical. The first of these is the one to which I personally have devoted the most work: the problem of developing a representation of spatial information suitable for commonsense physical reasoning.

For almost all purposes, we may take the ontology of space to be Euclidean space. This is, indeed, one of the great advantages of the competence-theory / model-based approach. What the cognitive or the naive theory (or theories) of space actually *is* is a very difficult question, but it is almost indisputable that Euclidean space is sufficient. To be precise, with the possible exception of one small category, I don't know of any inferences in commonsense physical reasoning where the Euclidean model of space is demonstrably inadequate, and I don't know of any commonsense physical inferences involving any substantive spatial reasoning[16] that can be characterized much more easily in some alternative model of space than in Euclidean space. (In particular, as far as I can tell, all the inferences considered in [Fleck, 96] can be characterized in a Euclidean model, with a suitable statement of the physical axioms.) The only exception I know of is order of magnitude reasoning, which is sometimes easier to do in a model of space based on the non-standard real line with infinitesimals.

Choosing Euclidean space as the ontology likewise gives us a substantial head start on the representation problem, as we can choose concepts and primitives from the vast array supplied by classical geometry. Commonsense physical reasoning does, indeed, require many different categories of spatial concepts. In our original scenarios, we encounter the following types of spatial relations:

- In all these scenarios, distinct objects do not overlap, a topological relation.
- In scenario 1, the string must be looped around the stake and the plant.[17] This is a topological relation between the string, the region consisting of the stake together with an infinite cylinder going up and down from the stake, and the region consisting of the plant together with an infinite cylinder going up and down from the plant. (It is not a topological relation between the objects involved as, topologically, the three objects are separate.)
- The string has to be long enough to loop around the stake and the plant. This is a relation between the length of a curve (the string), the distance between the plants, and the circumferences of the plant and stake at the point where they meet the string.
- The stake is stuck in a hole in the ground. That is, let G_0 be the original shape of the ground, let G_1 be the shape of the ground after the stake has been placed, and let S be the region occupied by the stake. Then (1) $G_1 = G_0 - S$; (2) $S \cap G_0$ includes one end of the stake and a cylindrical section of the stake.
- An unstaked plant will break if it bends too much. It is not clear how this should be best characterized; perhaps as a constraint on the maximum curvature of the central axis of the stalk of the plant.
- At the end of scenario 3, there is cold water inside the baby bottle; that is, topologically inside the region defined by the union of the bottle with the hole in the nipple.
- In scenario 3, the volume of water in the pot of ice-water is greater than the volume of the baby bottle.

[16] There are inferences involving trivial spatial reasoning that are more easily characterized in simpler spatial models. For instance, in the blocks world, space can be taken to be a finite set of stacks.

[17] I am reliably informed [B.I. Davis, personal communication] that the string should be looped in a figure-eight, but I do not know what the physical reason for this is.

- There is a single small hole connecting the inside of the baby bottle to the outside.
- The baby bottle will fit in both pots below the top of the water. That is, for each pot, there is a rigid transformation of the shape of the bottle into the region occupied by the water in the pot.

Thus, these three scenarios involve at least the following spatial concepts:

- Metric dimension
- Length of a curve
- Volume of a region
- Set operations: Union, intersect, set-difference, subset, disjoint.
- Large-scale topological concepts: topological containment, looping around, etc.
- Rigid transformation
- Vertical vs. horizontal.
- Curvature.

Other commonsense inferences require further spatial concepts, including

- Tangent of a curve. A good example is NEWTON [de Kleer, 77], where the behavior of a roller coaster on a track is characterized in terms of the signs of the slope and curvature of the track.
- Relative position and relative orientation. E.g. Object A is between B and C. The upper right arm, in a particular configuration, points downward, forward, and out from the body.
- Angular dimension. E.g. the faces meeting at the cutting edge of blade must form a small angle.
- Continuous symmetry. E.g. A circular wheel can be spun in place. A helical screw can be drawn out of a hole.
- Strictly repeated structure. E.g. the arrangement of links in a chain, or of bricks in a wall.
- Collections of related but variable shapes in non-fixed relations. E.g. characterizing a pile of sand, a stack of hay, the limbs of a tree

Further, it seems plausible to suppose that a spatial language should have the capacity to describe exact shapes, parametrized exact shapes, and the notion of one shape approximating another. A mechanical part, for instance, has an ideal shape; the actual part lies within some tolerance of that ideal ([Requicha, 83], [Joskowicz et al., 96], [Davis, 96]). Different notions of shape approximation may be appropriate to different kinds of inference. For example, to deduce that the actual shape will fit inside a space that comfortably accommodates the ideal shape, it suffices to bound the "distance" (strictly speaking, the Hausdorff distance) between the two shapes. To ensure that the actual shape will move as smoothly as the ideal shape, it is necessary to further constrain the surface tangents of the actual shape to approximate those of the ideal shape.

The above predicates are all static. We additionally need characterizations of change and motion. These have been less studied than static spatial predicates. The following seem like plausible candidates.

- A temporal language applied to spatial relations, construed as fluents. E.g. The baby bottle is at one time in the cold water and at another time in the hot water. The bottle remains in the cold water for an interval of a few minutes duration.

- Linear and angular velocity and acceleration. The egg to be cracked must be brought down to the bowl fast but not too fast. A top spins quickly about its axis. An object moving up hill slows down [de Kleer, 1977].

- Topological characterizations of motion. The water flows *through* the hole in the nipple.

- Characterizing fields of flow. E.g. the water draining down a sink forms a whirlpool.

- Constraints on change of shape. E.g. the shape of a quantity of water changes under the Hausdorff metric.

This may start to look like a mere *omnium gatherum* of every geometric property I can think of. To counter such a perception, let me list some properties that are important mathematically but that I would guess will *not* be needed in commonsense reasoning: Affine and projective transformations; third and higher derivatives; higher-order moments (beyond the moment of inertia); Fourier coefficients; zero-measure, everywhere dense sets; small-scale topology, such as the distinction between a closed and an open set.[18]

I think it is correct to say that most of the spatial concepts encountered in commonsense physical reasoning have a natural expression within standard geometrical terminology. There are certainly some spatial properties where the correct expression is doubtful; for example, in expressing the fact that a plant cannot be bent more than a certain amount, that might, as suggested above, be best characterized as a constraint on the curvature of the axis of the spine of the plant, or it might be best formalized some other way. There is probably no way to know until we have a complete commonsense theory of the bending and snapping of supple objects. There are also spatial properties where we do not know what the general representation should be; for instance, one can recognize a tree, but what does it mean to be shaped like a tree? On the other hand, there are only a few physical inferences that can be derived purely from the constraint that something is shaped like a tree, so it is not clear that this is an important fact to express for *physical* reasoning (visual recognition is another matter).

Of course, spatial properties that are easily characterized in standard geometric terms will be easier to analyze than those that are not. Therefore, the fact that thus far Euclidean geometry seems to be a sufficient ontology and that standard geometrical terminology seems to be a largely sufficient language is not really evidence that they will be so in the long run. The fact that the world is in fact well approximated by Euclidean geometry at all scales between the sub-atomic and the cosmological does not imply that the conceptual space need be Euclidean, any more than the fact that terrestrial physical objects are in fact made up of atoms implies that commonsense physical concepts can or should be given a semantics in terms of atomic structure. I do not see any way to decide this question positively until some large fraction of qualitative physics has in fact been formalized.

The real difficulty in spatial reasoning is computational. The tradeoff between expressivity and tractability is particularly grim in spatial reasoning; very limited geometric languages turn out to be wildly intractable or undecidable. The problem of finding a language in which all the spatial concepts we need are expressible together with an inference engine that can efficiently carry out all the inferences we need is therefore very severe and will require a very careful examination of

[18] This distinction is useful to the theory designer, who may want to specify that all objects occupy regular sets or some such other constraint. I suspect, however, that it is not needed as an explicit concept in the representation language; that is, the representation language will not need predicates like "open(R)" or "closed(R)" or operators such as "interior(R)" or "closure(R)".

the space of inferences. (Again, as discussed in section 9.1 it is a pity that our notion of "all the inferences we need" is not better defined.)

Five general directions of research suggest themselves:

1. Start with a sufficiently expressive language, and winnow it down toward tractability.

   We may begin on this path by observing that, if we are permitted to use the full power of first-order logic and to quantify over spatial regions, then the result is a language with much more expressivity than you need. For instance, if you take the first-order language where variables range over points in space and regions in space, and include the two predicates

   closer(X, Y, Z) — point X is closer to point Y than to point Z.
   member(X, S) — point X is an element of set S

   then you get all of second-order analysis, which includes all the static concepts enumerated above, both the ones we need and the ones we don't, and, indeed, practically every other relation on spatial regions defined in the mathematical literature. The dynamic relations can be all added by adding metric time and histories.

   This suggests that we should vastly restrict the logical form of the language; for example, restrict the language to constraints (atomic ground formulas). A proposal for a richly expressive language of constraints, with lots of primitive predicates, is advanced in [Davis, 1996]. However, this does not bring us anywhere near close enough to tractability. It is easily shown that many very small subsets of this language are intractable; some are undecidable. The next step from this end would presumably be to study the effects of restricting the structure of the constraint network on tractability and expressivity. This technique has been extensively studied for discrete constraints, but very little is known about it in the context of spatial constraints.

2. Start with a limited language of very general primitives and gradually add more types of information. This approach has been followed by A.G. Cohn and his associates, who began by studying basic topological relations [Randell and Cohn, 89], and have gradually expanded the language by adding such properties as convexity [Randell, Cui, and Cohn, 92]. This approach has been probably the most productive in terms of interesting and suggestive qualitative spatial languages. The difficulty in this approach, aside from the inevitable intractability cliff, is the fact that the basic primitives don't allow you to express anything of any interest. One has to look very hard to come up with plausible examples of physical inference expressible in the RCC language.

   Also falling into this category are the use of the sign calculus, as in NEWTON [de Kleer, 77], FROB [Forbus, 80], and so on, and the use of the cross-product of Allen's [1983] interval relations, as suggested in [Lenat and Guha, 90] and elsewhere. These, however, are all of limited applicability, since they all depend critically on a standard coordinate system, and are usefully expressive only of relations that line up neatly along that system.

3. Start with a language of precise spatial information — that is, a diagrammatic representation [Glasgow, Narayanan, and Chandrasekaran, 95] — and then add primitives for replacing numeric values with constrained parameters [Brooks, 81], [McDermott and Davis, 84] and for stating that the true shape is approximation of the ideal shape [Requicha, 83], [Davis, 86], [Nielssen, 88] [Joskowicz and Sachs, 96], [Davis, 96]. Ideally, this kind of approach should preserve the intuitive appeal and some of the computational simplicity of diagrammatic representations, while allowing a useful range of partial information to be expressed.

   The difficulties here are, first, that many common types of partial information are not naturally expressed in these forms; second, that parameterized shape and shape approximation leads to hard computational problems; third, that showing that physical properties are preserved over the space of parameter values and approximating shapes tends to be very difficult, particularly if uncertainty involved is not extremely small as compared to the dimensions of the problem.

4. Develop a language of spatial relation based on spatial expressions in natural language [Bloom et al., 1996]. This approach gives many interesting results. In particular spatial words turn out to be not purely geometric, but depend strongly on physical relations, such as the support relation between objects. The chief limitation of this approach, for our purposes, is that the spatial relations expressible in non-technical language are not sufficient to support physical reasoning. In particular, the language of shape is very attenuated.

5. Find the spatial relations and properties that are used in rules of physical inference. This gives a spatial language tuned to our particular needs. The problem here is that there is no guarantee that the spatial properties and relations that occur naturally in physical rules are commonsensically important. For instance, a string hangs in a catenary, but is the catenary an important geometric primitive?

11 The Extended Prediction Problem

In most implemented physical reasoning systems,[19], temporal reasoning is carried out in sequential steps of time, calculating all expressible aspects of each distinguishable state in order of occurrence. (What aspect of states are expressible and what states are distinguishable depends on the language of states used.) Often such a complete account is difficult to attain and useless when attained. For example, consider trying to predict the behavior of a block dropped on the ground. The sequence of states after the first collision and before setting down are typically very complicated and very unstable; the block bounces, twists, and slides around in a manner whose details depend very delicately on the exact shapes, material characteristics, and initial conditions. For most purposes, they are also entirely unimportant; the behavior of the block can be adequately summarized in the statement that it fall, hits the ground, rattles around briefly at a low height, and comes to rest not far from where it first hit. Another example, due to Drew McDermott [personal communication]: In July, you observe that a couple of sticks of dynamite have been stored in the fireplace. You want to be able to infer that these should be removed before winter, without attempting to predict everything that happens in between. Failure to address this kind of inference is a common characteristic of theories that are metaphysically adequate but not epistemically adequate [McCarthy and Hayes, 69].

12 Multiple scales

Physical reasoning often involves combine reasoning about entities and parameters of widely different magnitudes. Some examples:

- The first leg of a cross country trip of three thousand miles and two weeks may be to spend five seconds backing six feet out of a parking spot. The ratio in time here is $2 \cdot 10^5 :: 1$; the ratio in space is $3 \cdot 10^6 :: 1$.
- The Challenger rocket, which was 150 feet tall, exploded because of its O-rings failed to expand by 1/200 inch. The ratio is $4 \cdot 10^5 :: 1$.
- A large organism may become sick or die because of a virus. The ratio of the mass of a man to the mass of a virus is roughly $10^{22} :: 1$.

[19] [Williams, 86] is an important exception.

- The shining of the sun is explained in terms of nuclear reactions. The ratio of the mass of the sun to the mass of the proton is $10^{57} :: 1$.

Automating reasoning in such domains requires dealing with issues that do not arise, at least not with the same force, in domains where everything important is more or less the same size. Most obviously, diagrammatic methods and complete simulation cannot be directly applied. It is not sensible to map out the whole of a cross-country trip to a precision of six feet and to schedule it to a precision of five seconds. It is impossible to simulate the sun by tracking 10^{57} nuclear particles.[20]

The problem of reasoning over multiple scales has a number of different aspects:

- Order of magnitude reasoning over scalars; the calculus of very small and very large numbers. The algebra of this theory is quite well understood (e.g. [Raiman, 86], [Dague, 93]). This theory justifies inferences like "If A is very much bigger than B, then $A+B$ is very close to A.' The differential and integral calculus in this theory have been much less developed in the AI literature; little has been done beyond [Weld, 87] and [Davis, 89]. For instance, there is no implemented system capable of carrying out the following inference: A small change to the parameters of a damped harmonic oscillator results in a small change in the period of the oscillator.
- Order of magnitude geometric reasoning. This includes such inferences as:
  - If the distance from **p** to **q** is much greater than the distance from **q** to **r**, then the angle $\angle\mathbf{rpq}$ is much less than 1.
  - Let **O** be an obstacle, and let **p** and **q** be two points outside of **O**. Suppose that the distance from **p** to **q** is much greater than the diameter of **O**. Then the length of shortest path from **p** to **q** that avoids **O** is only negligibly greater than the distance from **p** to **q**.
  - If **p** and **q** are points on a curve **C**, whose curvature is never greater than κ, and $\kappa \mid \mathbf{p}-\mathbf{q} \mid \ll 1$, then the angle between the line **pq** and the tangent to **C** at **p** is much less than 1.
- Temporal, spatial, and physical hierarchies. Developing representations that integrate crude high-level representations with finer detail where necessary, decomposing either temporally [Allen, 83], [Koomen, 89], geometrically [Glasgow and Papadias, 95], or structurally [Sussman and Steele, 80]. Much remains unknown about best ways to characterize relations between levels of the hierarchy and how to derive reliable inferences from the higher levels of the hierarchy that do not have to be checked against the fully detailed description.
- Ensembles. How the behavior of a large collection of similar objects relates to the behavior of the individual objects and the structure of the collection. For instance, how the behavior of a chain relates to its structure and the behavior of the links; how the behavior of a heap of sand relates to the individual grains.

13 Plausible Inference

Many of the rules used in physical reasoning are not sound, deductive rules but rather rules of plausible inference. Previously in this paper, for instance, we have seen the following plausible rules:

[20] It may, of course, be possible to use several different diagrams each at a single scale, but the problems of relating and combining information across diagrams at different scales are not substantially easier than the problem of reasoning symbolically at different scales. A collection of diagrams does not not have the advantages generally attributed to diagrammatic representations [Davis, 91].

> The thinner you roll the dough, the more cookies you get.
>
> If you cut through an object anywhere near the center, in all likelihood you will destroy its functionality.

Following in the spirit of the microworlds methodology, a natural approach would be to define, for a given microworld, a prior probability distribution on the models (in the strict sense) of that microworld. The conditional probabilities are then defined in terms of the prior probabilities. More precisely, since there are generally uncountably many models, we need a measure over significant classes of models. Indeed, in order to accommodate conditions that reduce the dimensionality of the space of models, we need such a measure for each significant subspace of models. For commonsense reasoning, we probably do not need a measure that is in any particular sense optimal, but we do need one that is well-defined.

It is certainly not obvious how to do this in general, but it may be, in fact, be reasonably easy to attain, at least for simple microworlds. For example, in the kinematics rigid solid objects, for a fixed number of objects, we can take the motions and shapes of objects in an initial position to be independent parameters, and then excluding from the sample space models where the non-overlapping condition is violated. For the object motions, there are standard random distributions that we can use. The more difficult issue is the random variation over object shapes. I don't know of any standard probability distributions over the sample space of geometric regions, but it may be possible to define one.

In cases like these, defining the prior probability distribution determines all the ontological inferences raised by the plausible inference. The only extension to the representation needed is the sentential probability operator. What the computational issues look like, we cannot begin to say.

Of course, an approach like this will work much better for inferences about properties that are easily characterized in a model, such as "The thinner the dough, the more cookies," than for more nebulous properties, such as "Cutting through an object near the center will probably destroy its functionality." Still less will it work for using plausible inference to guess the properties of a microworld. For instance, consider the following argument by analogy:

> A chain is like a string. Therefore, since cutting a string in half gives two strings, probably cutting a chain in half will give two chains.

It is hard to imagine what sample space the "probably" there refers to or, more generally, to see any way of dealing with this in a model-based theory.[21]

14 Multiple models, approximation, and abstraction

A key aspect of effectively solving physical problem at either the commonsensical or the expert level involves choosing the appropriate physical model, approximating a complex situation or theory by a simpler one, or abstracting out the key element of the given problem and ignoring the inessential elements.

[21] The orthodox Bayesian approach, of course, gives a simple prescription: Present yourself with a series of wagers that cutting a chain in half gives two chains, and define the probability based on the point where you feel the wager is fair. In fact, if you adopt any other procedure, then you are provably opening yourself to losing all your money in Dutch games to one of the wandering gamblers who prey on non-Bayesians. Somehow, though, I don't find that this gets me very far in my difficulties.

The use of multiple models in physical reasoning is so natural and ubiquitous that it is often supposed that the knowledge base of physical knowledge as a whole is best analyzed in terms of a structure of models and rules for operating over these models [Addanki et al., 89]. However, it seems to me that this view is seriously at variance with the goals of a declarative representation or a knowledge-based analysis. The argument is as follows: If you have multiple models, then inevitably there will be problems that you end up using both models together in a single problem, because the two models will apply at different times in the problem or at different places, or to different objects, or at different scales of granularity. At this point, you are making inferences from inconsistent axioms, which is a meaningless operation. To avoid getting into trouble, and to insure that you don't end up inferring nonsense, you will have to limit inference in some ways. It will be difficult to characterize these limitations, difficult to justify them, and difficult to be sure that you have succeeded in blocking all destructive interactions between the conflicting models.

A better approach, it seems to me, is to look for an single overall theory $\mathcal{T}$ that in principle supports the entire class of inferences we wish to achieve. The use of alternative models and the rules to choose between these or to combine these can then be justified in terms of $\mathcal{T}$ as approximations or heuristics that are (likely to be) valid for a given problem and simplify computation.

Techniques for abstraction and approximation have been the subject of considerable study recently. It seems to me, however, that there are two erroneous assumptions that run through much of this work.

First, it is often assumed that approximating a complex theory by a simpler one always simplifies calculation, though at the cost of reducing accuracy [Weld, 92]. This is not the case: Depending on the problem, simpler theories may actually be harder to compute with. Some examples:

- Zero friction. If the coefficient of friction in figure 7 is 1/4, then the system is static; if it is zero, then the system will evolve in a complex way.
- Elastic collisions. If the coefficient of restitution is less than 1/2, then the system in figure 8 will, within finite time, settle to an equilibrium state with the ball at the bottom. If collisions are fully elastic, then the long term behavior is more complicated. (Figures 7 and 8 look rather specialized, but in general the world would be much more complicated and difficult to handle if there were no friction and all collisions were perfectly elastic.)
- Non-relativistic motion. Consider the following problem: An electron has kinetic energy of 1 Gev. How long does it take to traverse a foot? Relativistically, this does not even require the back of an envelope: The stated kinetic energy is much greater than the rest mass of the electron, so the electron is moving at almost the speed of light = 1 foot per nanosecond. Hence 1 nanosecond is the answer. The non-relativistic answer is not only wrong, but requires more work to find.
- Uniform gravitational field. A projectile is fired out from the earth at 1000 miles per second. What is its state ninety-six hours hence? The correct answer, with an inverse square gravitational law, is that, since gravity is negligible after the first few minutes, the projectile is very slightly less than 345,600,000 miles away and continuing to move outward at very slightly less than 1000 miles per second. The answer calculated with a uniform gravitational field, ignoring the motion of the earth, is that six hours ago it crashed back into the earth, with consequences that are not easy to predict in detail. If the motion of the earth is considered, then it is somewhere fairly near earth; where, exactly, would take a good bit of work to determine. (The correct answer can be found efficiently if the gravitational law is approximated as "zero gravity".)

(The last two examples above may seem contrived. However, it is not clear what "natural" space of problems is being considered here. In the ordinary course of running a household or designing a

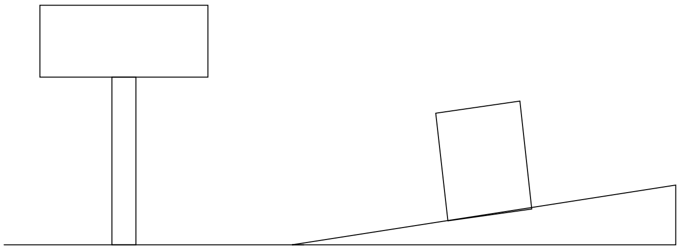


Figure 7: A simple system with friction

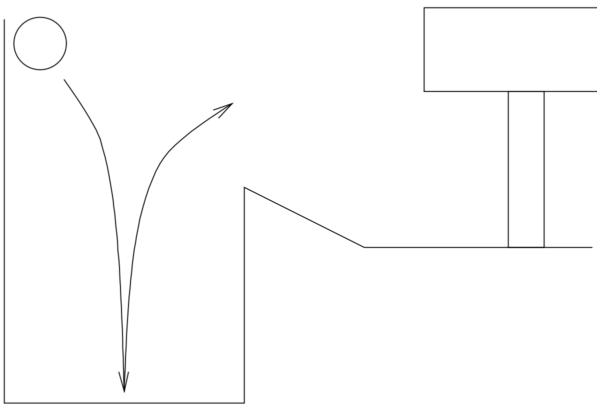


Figure 8: A simple system with inelastic collisions

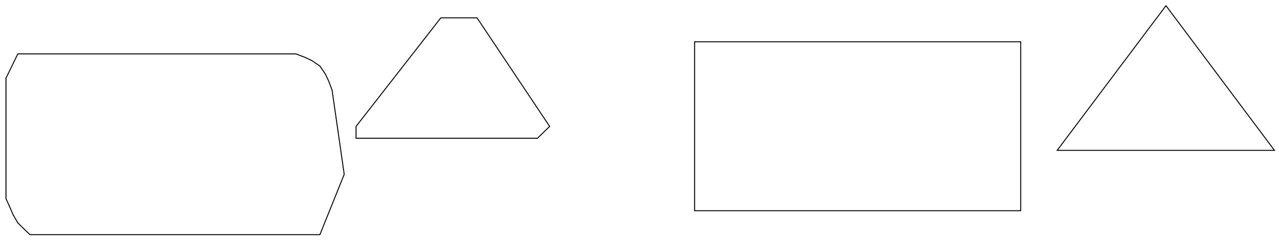


Figure 9: Shape approximation

mechanism, a physical reasoner would never have to consider either the relativistic correction or the correction for non-uniform gravitational field — they are too small to make any kind of a difference. Therefore, if a reasoner does include such theories as alternates worth considering, the problem space being considered must include some problems with relativistic speeds / heights comparable to the radius of the earth.)

Certainly, there are pairs of theories where the more idealized theory is essentially always simpler to use. One would have to work hard to contrive a problems where it was easier to assume temperature-dependent resistivity than constant resistivity, or where it was easier to use to use the Van der Waals equation than to use the ideal gas law.[22] But this simple relation is not as ubiquitous as Weld's discussion suggests.

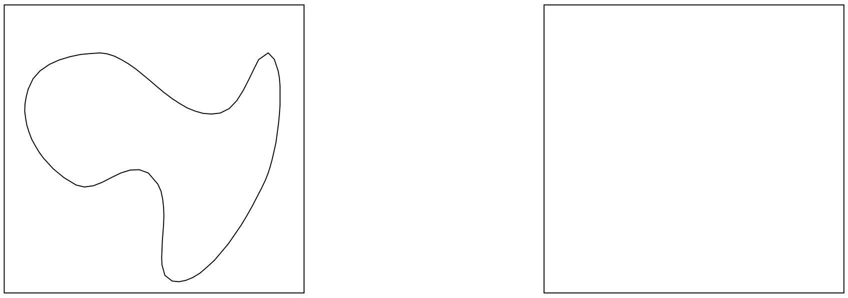
The second error is to view abstraction as primarily a relation between theories ([Giunchiglia and Walsh, 92], [Nayak and Levy, 94]). In [Davis, 95] I did a study of abstraction methods that could be applied to problems in the kinematic theory of rigid solid objects (KRSO). I found that there were abstractions of all the following types:

I. Approximation of geometric detail. For instance, in figure 9, the constraints on the motions of the objects in (B) are almost the same as in (A), and much easier to compute. In figure 10, the internal cavity can be ignored, since it cannot affect the behavior of objects outside the rectangle.

II. Approximations of structure. A collection of objects is abstracted as a smaller collection of objects. For instance, in figure 11, if you are solving a problem that refers only to objects o1, o2, o3, and o4, it is possible to ignore the existence of o5 and o6. In figure 12, you can coalesce the four objects into a single object, since they are tightly bound together.

III. Abstraction of a problem in KRSO by a problem in a simpler theory. For instance, the standard abstraction of the 15-puzzle (Figure 13) abstract the continuous problem of moving square pieces to an discrete problem of swapping the blank entry in an array with a neighboring entry. Note that finding this abstraction requires a great deal of reasoning: You have to figure that all positions where the squares are not aligned in rows and columns can be ignored, and that the physical option of moving two squares at a time in the row/column with the blank can likewise be ignored.

[22]It is always possible to contrive an example where theory A is simpler than theory B as follows: Develop an experiment that tests between A and B. Attach an immensely complicated Rube Goldberg device to the output of the test in such a way that the device is triggered only if B is satisfied. Then prediction using B, which involves analyzing the device, is more complicated than prediction using A, which involves only recognizing that the device is not triggered.



The internal cavity can be ignored.

Figure 10: Abstracting away inaccessible boundary parts

IV. Abstraction of a problem in KRSO by a simpler problem in a more complex theory. For instance, a long chain linking two rigid objects can be approximated kinematically as a string linking those objects. The theory of strings and solid objects is more complicated than just the theory of solid objects; however, the simplicity gained by reducing the many links of the chain to the single string makes up for it.

Based on these examples and on the considerations at the beginning of this section, I suspect that abstraction and approximation are best analyzed as techniques of problem reduction rather than as relations between physical theories.

15 The next stage

Despite all these difficulties and objections, I find our original scenarios — the staked plant, the cookie dough, the baby bottles, and a myriad similar situations — too fascinating and compelling to abandon. I still feel that it is wise to begin by developing representations for a knowledge-level analysis, and that the method of microworlds is the most promising approach that we have. The main task now, therefore, is to develop more and richer microworlds.

As discussed in section 9.5 , we can expect the next generation of microworlds will be more difficult in every respect than those we have already seen. If we look at microworlds such as the dynamics of cutting, we expect to find that microworlds will be more complex and narrower; that reasoning will rely more on plausible inference; that the spatial component of reasoning will be both more complex and less clearly defined; and that immediate connections to useful applications will become fewer. But, if we have patience enough to stick with it, we should eventually have a remarkable theory.

References

James Allen (1983). "Maintaining Knowledge about Temporal Intervals." *Comm. ACM* vol. 23, pp. 832-843.

Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett (eds.) (1996). *Language and*

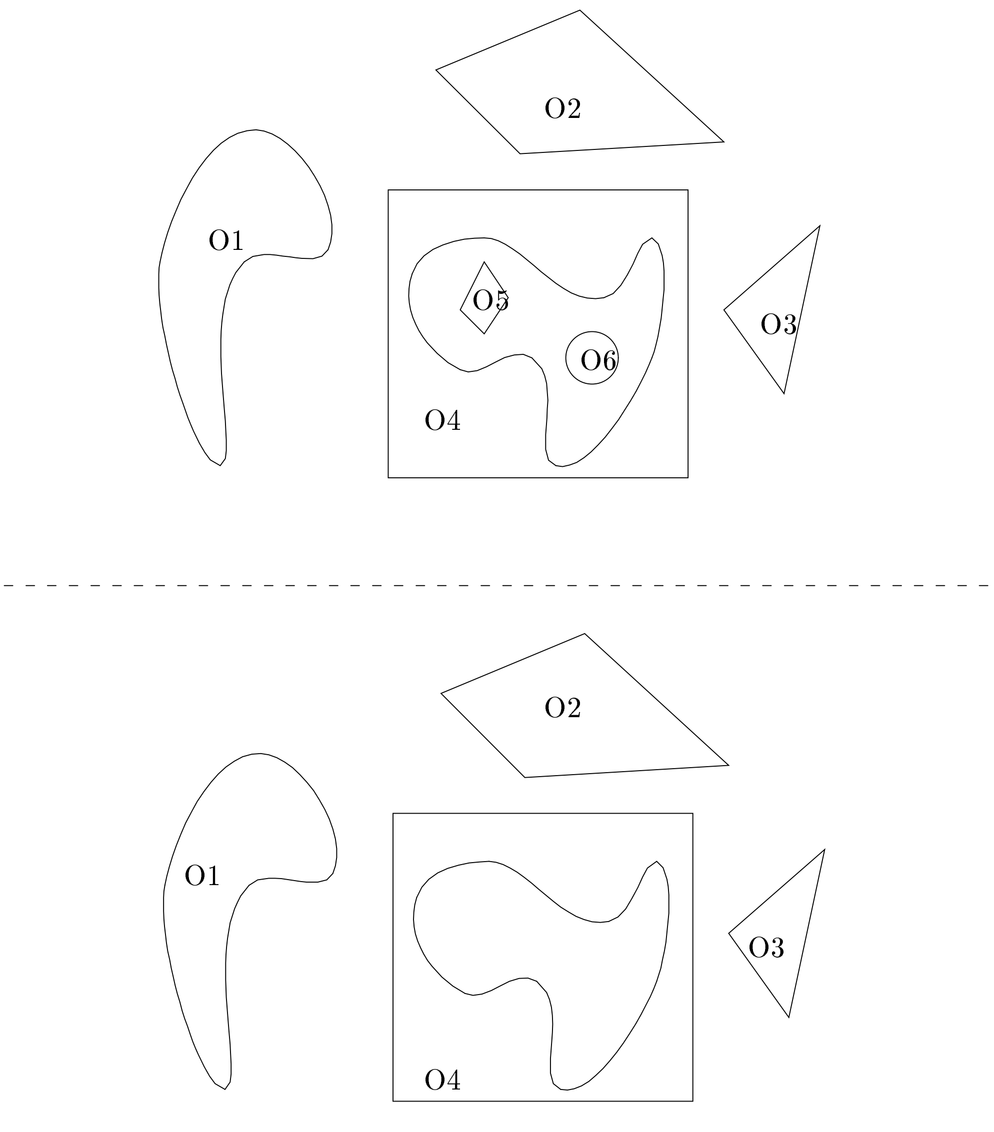


Figure 11: Ignoring inaccessible objects

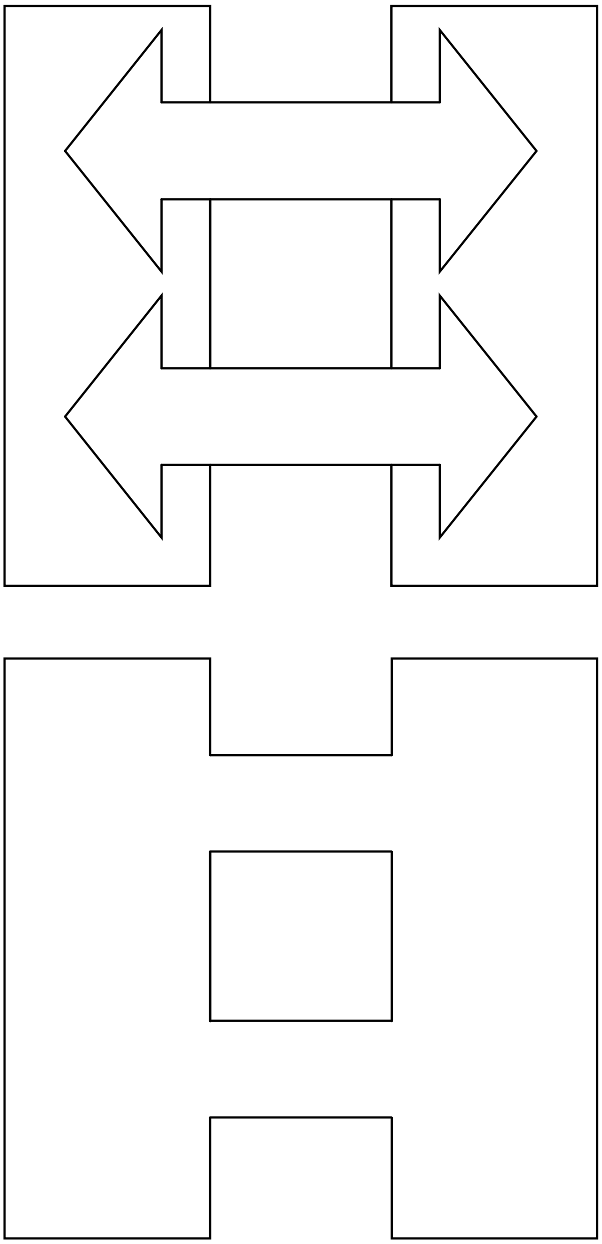


Figure 12: Coalescing bound objects

B	M	O	G
A	J		E
N	C	I	H
L	D	K	F

$$\begin{bmatrix} B & M & O & G \\ A & J & \_ & E \\ N & C & I & H \\ L & D & K & F \end{bmatrix}$$

Figure 13: The fifteen puzzle and its abstraction

Space, MIT Press.

Rodney Brooks (1981). "Symbolic reasoning among 3-D models and 2-D images." *Artificial Intelligence,* vol. 15, pp. 285–348.

Eugene Charniak and Drew McDermott (1985) *Introduction to Artificial Intelligence.* Addison-Wesley.

Philippe Dague (1993). "Symbolic reasoning with relative orders of magnitude." *Proc. IJCAI-93,* pp. 1509-1514.

Ernest Davis (1986). *Representing and Acquiring Geographic Knowledge.* Pitman Press, London.

Ernest Davis (1991). "Lucid Representations." Tech. Rep. 565, NYU Comp. Sci. Dept., June 1991.

Ernest Davis (1993). "The Kinematics of Cutting Solid Objects." *Annals of Mathematics and Artificial Intelligence,* vol. 9, no. 3,4, 1993, pp. 253-305.

Ernest Davis (1995a) "Approximations of Shape and Configuration Space." NYU Computer Science Tech. Report #703.

Ernest Davis (1995b). "Approximation and Abstraction in Solid Object Kinematics." NYU Computer Science Tech. Report #706.

Ernest Davis (1996). "A Highly Expressive Language of Spatial Constraints." NYU Computer Science Tech. Report #714.

Johan de Kleer (1977). "Multiple Representations of Knowledge in a Mechanics Problem Solver." *Proc. IJCAI-77.*

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

Margaret Fleck (1996). "The topology of boundaries." *Artificial Intelligence,* vol. 80, pp. 1-27.

Boi Faltings (1987). "Qualitative Kinematics in Mechanisms." *IJCAI* pp. 436-442.

Kenneth Forbus (1980). "Spatial and Qualitative Aspects of Reasoning about Motion." *Proc. AAAI-80.*

Kenneth Forbus (1985). "Qualitative Process Theory." In D. Bobrow (ed.) *Qualitative Reasoning about Physical Systems,* MIT Press.

Fausto Giunchiglia and Toby Walsh, "A theory of abstraction," *Artificial Intelligence*, vol. 57, pp. 323-389, 1992.

Janice Glasgow, N.H. Narayanan, and B. Chandrasekaran (1995). *Diagrammatic Reasoning,* MIT Press.

Janice Glasgow and D. Papadias (1995). "Computational Imagery." In Janice Glasgow, N.H. Narayanan, and B. Chandrasekaran (eds.) *Diagrammatic Reasoning,* MIT Press.

Joseph Halpern and Moshe Vardi (1991). "Model Checking vs. Theorem Proving: A Manifesto." *Second Intl. Conference on Knowledge Representation and Reasoning,* pp. 325-334.

Patrick Hayes (1977). "In Defense of Logic." *Proc. IJCAI-77*, pp. 559-565.

Patrick Hayes (1978). "The Naive Physics Manifesto" Working paper, Department of Computer Science, Univeristy of Essex.

Patrick Hayes (1979). "The Naive Physics Manifesto." In D. Michie (ed.) *Expert Systems in the*

Microelectronic Age. Edinburgh University Press.

Patrick Hayes (1985). "The Second Naive Physics Manifesto" In J. Hobbs and R. Moore, (eds.) *Formal Theories of the Commonsense World.* Ablex Pubs., Norwood, N.J.

Patrick Hayes (1985b). "Naive Physics 1: Ontology for Liquids." In J. Hobbs and R. Moore, (eds.) *Formal Theories of the Commonsense World.* Ablex Pubs., Norwood, N.J.

Yumi Iwasaki (ed.) *Qualitative Reasoning: When Precise Data Isn't What You Need.* Special Issue of *IEEE Expert: Intelligent Systems and their Applications,* vol. 12 no. 3.

Leo Joskowicz and Elisha Sacks, (1991). "Computational Kinematics," *Artificial Intelligence,* vol. 51, pp. 381-416.

Leo Joskowicz, Elisha Sacks, and Vijay Srinivasan, "Functional Kinematic Tolerancing," *Computer-Aided Design,* 1996, to appear.

J. Koomen, (1989). "Localizing Temporal Constraint Propagation." *Proc. 1st Int. Conf. on Principles of Knowledge Representation and Reasoning,* Morgan Kaufmann.

Robert Kowalski (1979). "Algorithm = logic + control." *Comm. ACM*, vol. 22, pp. 424-436.

Benjamin Kuipers (1986). "Qualitative Simulation." *Artificial Intelligence,* vol. 29, pp. 289-338.

Douglas Lenat, M. Prakash, and M. Shepherd (1986). "Cyc: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks." *AI Magazine,* vol. 6 pp. 65-85.

Douglas Lenat and R.V. Guha (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project.* Addison-Wesley.

Douglas Lenat and R.V. Guha (1993). "Re: CycLing paper reviews." *Artificial Intelligence,* vol. 61, pp. 149-174.

Q. Ji and M. Marefat (1997). "Machine Interpretation of CAD Data for Manufacturing Applications," *ACM Computing Surveys,* Sept. 1997.

Matthew Mason, "On the Scope of Quasi-Static Pushing", *Proc. 1985 Third Int. Symp. on Robotics Research,* MIT Press, 1985.

John McCarthy (1968). "Programs with Common Sense." In M. Minsky (ed.) *Semantic Information Processing.* MIT Press, pp. 403-418.

John McCarthy and Patrick Hayes (1969). "Some Philosophical Problems from the Standpoint of Artificial Intelligence." In B. Metzler and D. Michie (eds.) *Machine Intelligence 4*, Edinburgh University Press, pp. 463-502.

Michael McCloskey (1983). "Naive Theories of Motion". In D. Gentner and A. Stevens (eds.) *Mental Models*, Lawrence Erlbaum Assocs., Hillsdale, NJ.

Drew McDermott (1978). "Tarskian Semantics, or No Notation without Denotation!" *Cognitive Science,* vol. 2, pp. 277-282.

Drew McDermott (1987). "A Critique of Pure Reason." *Computational Intelligence,* vol. 3, pp. 151-160.

Drew McDermott and Ernest Davis (1984). "Planning Routes through Uncertain Territory." *Artificial Intelligence.* vol. 22, pp. 107–156.

Robert Moore (1982). "The Role of Logic in Knowledge Representations and Commonsense Reasoning." *Proc. AAAI* pp. 428-433.

P. Pandurang Nayak, "Causal Approximations" *Artificial Intelligence* vol. 70, 1994, pp. 277-334.

P. Pandurang Nayak and Alon Levy: "A Semantic Theory of Abstractions: Preliminary Report," *Workshop on Theory Reformulation and Abstraction,* 1994.

Alan Newell (1980). "The Knowledge Level." *AI Magazine,* vol. 2 no. 2, pp. 1-20.

Paul Nielssen, (1988). "A Qualitative Approach to Mechanical Constraint," *Proc. AAAI-88*, pp. 270-274.

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

D. A. Randell and A. G. Cohn (1989). "Modelling Topological and Metrical Properties in Physical Processes," *First Intl. Conf. on Principles of Knowledge Representation and Reasoning.* pp. 357-367.

D. A. Randell, Z. Cui, and A. G. Cohn (1992). "A Spatial Logic Based on Regions and Connection," *Third Intl. Conf. on Principles of Knowledge Representation and Reasoning.* pp. 165-176.

Aristides A.G. Requicha (1983). "Towards a Theory of Geometric Tolerancing," *The International Journal of Robotics Research,* vol 2, no. 4, 1983, pp. 45-60.

Bassam Z. Shakhashiri (1985). *Chemical Demonstrations: A Handbook for Teachers of Chemistry.* University of Wisconsin Press.

James G. Schmolze (1986). "Physics for Robots." *Proc. AAAI-86,* pp. 44-50.

Gerry Sussman and Guy Steele (1980). "CONSTRAINTS: A Language for Expressing Almost Hierarchical Descriptions." *Artificial Intelligence,* vol. 14, pp. 1-40.

Mark S. Tuttle (1993). "Book Review of E. Davis *Representations of Commonsense Knowledge*, and D.B. Lenat and R.V. Guha, *Building Large Knowledge-Based Systems: Representations and Inference in the Cyc Project*," *Artificial Intelligence,* vol. 61, pp. 121-148.

Daniel Weld, "Reasoning about Model Accuracy," *Artificial Intelligence,* vol. 56, 1992, pp. 255-300.

Daniel Weld and Johan de Kleer (1989). *Readings in Qualitative Reasoning about Physical Systems.* Morgan Kaufmann, San Mateo, Cal.

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

Brian Williams (1986). "Doing Time: Putting Qualitative Reasoning on Firmer Ground," *Proc. AAAI-86*, pp. 105-113.