AI Magazine Volume 19 Number 4 (1998) (© AAAI)

The Naive Physics Perplex

Ernest Davis

"Common sense is a wild thing, savage, and beyond rules."
—G. K. Chesterton, *Charles Dickens: A Critical Study* (1906, p. 136)

■ The "Naive Physics Manifesto" of Pat Hayes (1978) proposes a large-scale project to develop 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 on 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 article, I compare the advantages and disadvantages of the two approaches.

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 might bend the plant. The fragile stem, bent too far, might snap, killing the plant. However, 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."

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. Because the plan, as given earlier 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 might be to encircle the plant with a metal frame.

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

The depth and power of our understanding seems to be most readily exhibited by this ability of exploring variants. Over a limited class of plans, explanations and execution sequences can be canned, or generated by, narrow special-purpose techniques. Moreover, the difficulties in writing an adaptable text generator or plan executor are mostly those of natural language

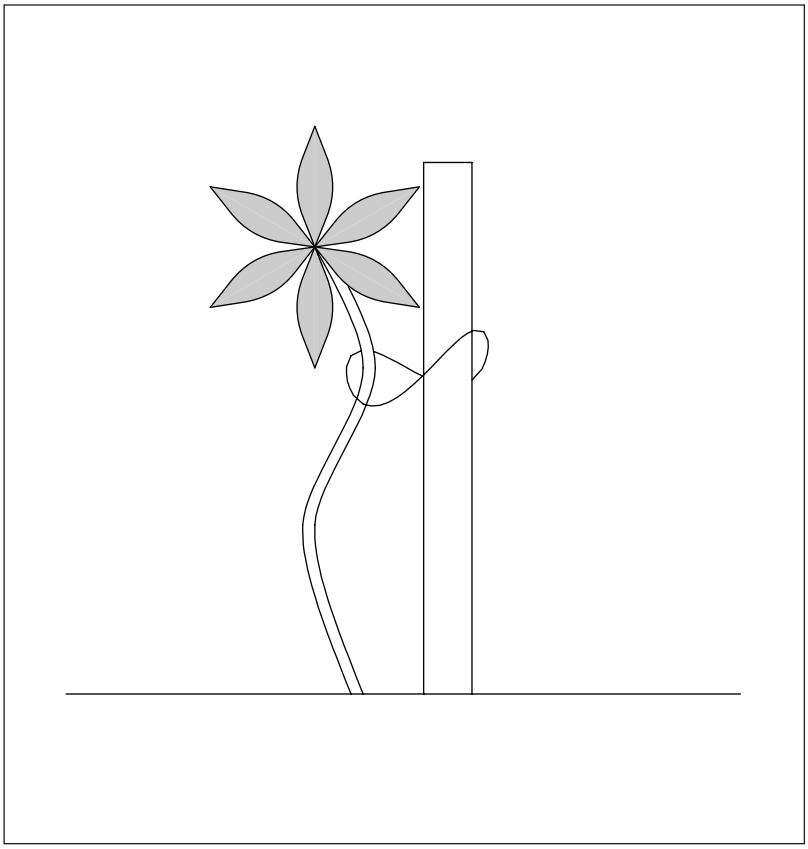

Figure 1. Staking a Plant.

and robotics, respectively; in practice, these issues swamp the problems of representation and reasoning. Adaptation and alternative application of plans certainly show understanding but might require an exceptional level of ingenuity. However, anyone who understands 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: 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, cut out cookie shapes with a cookie cutter, and put the separated cookies onto a cookie sheet and bake.

What happens if you do not flour the surface? What if you use too much flour? What if you do not roll out the dough but cut the cookies from the original mass? What if you roll out the dough but don't cut it? What if you cut the dough but don't separate the pieces? What happens if the surface is covered with sand or covered with sandpaper? What if the rolling pin has bumps or has cavities or is square? What 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? What 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? What if the cuts with the cutter overlap? Does the dough end up thinner or thicker if you exert more force on the rolling pin? What if you roll it out more times, or you roll the pin faster or slower? Do you get more or fewer cookies if the dough is rolled thinner or a larger cookie cutter is used? What if there is more dough? What if the cuts with the cutter are spread farther apart?

Scenario 3: The following experiment is described in Shakhashiri (1985) for estimating absolute zero using household objects. Prepare a pot of boiling water and a pot of ice water. Take an empty graduated baby bottle, complete with nipple attached, and submerge it (using tongs) in the boiling water. After a few minutes, when it has stopped bubbling, remove it and plunge it rapidly under 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 can be measured by holding the bottle right-side up; the final volume of the gas at 0° C can 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 is zero, one can find the value of absolute zero (figure 2).[1]

What would happen if the bottle is immersed only briefly in the hot water or only briefly in the cold water? What if it is laid on top of the pots of water rather than immersed in them? What 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? What if the bottle has an open end with no nipple or if the nipple has no hole? What if the bottle has other holes besides this nipple? What if you use containers with air at 100° C and 0° C rather than water? What if the quantity of ice water in

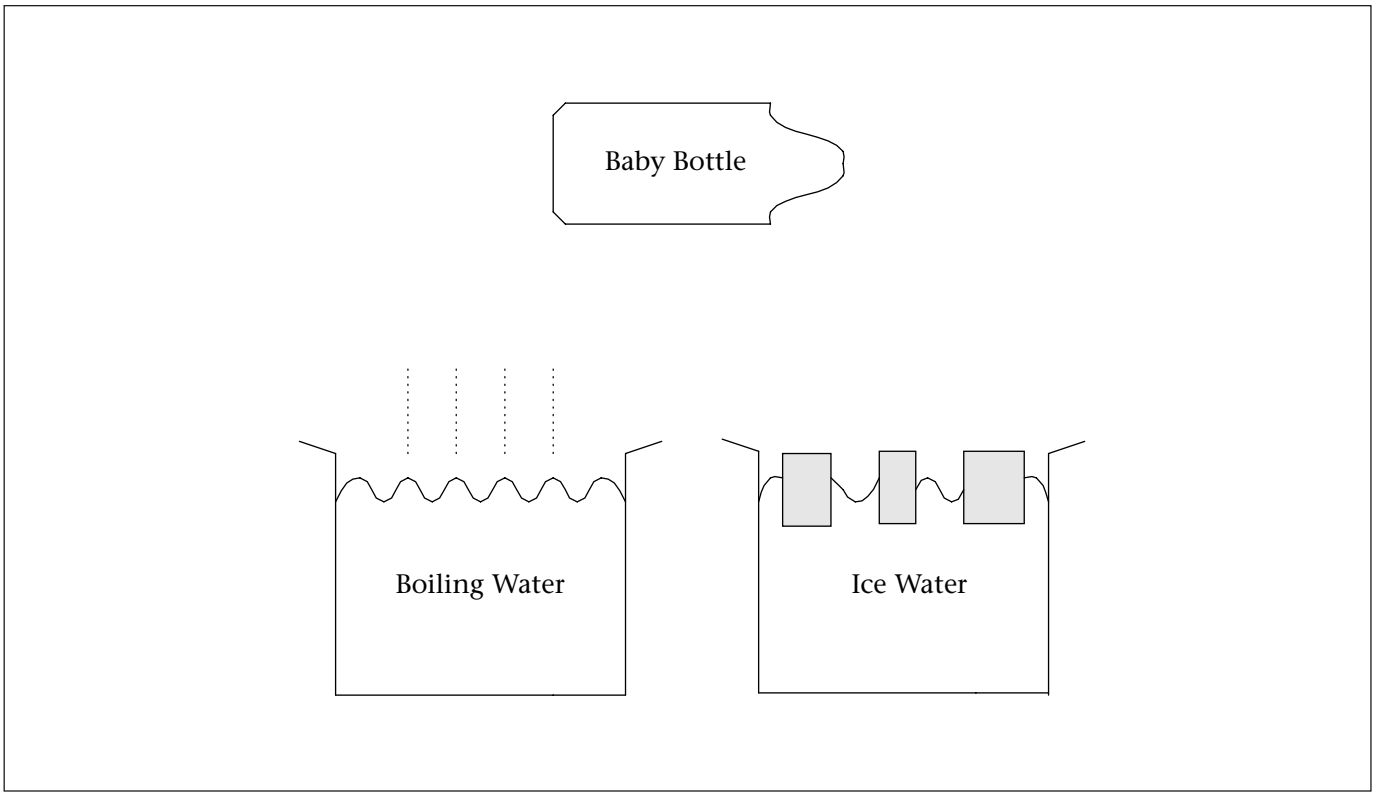

Figure 2. Determining Absolute Zero.

the second pot is very small or very large or if the quantity of hot water in the first pot is very small or very large? What if the bottle is coated with Styrofoam? What if the bottle is opaque? What 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."

Additional problems of this flavor in commonsense reasoning can be found in Miller and Morgenstern (1998).

Commonsense Physics

These three scenarios 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 gases, 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?" "What happens if the quantity of cold water is very small?" However, precise quantitative values are rare, and textbook-style equations are practically nonexistent, 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 string must encircle the stake and the plant. The bottle must not have holes other than the nipple and must be immersed in the water. However, no precise geometric descriptions are given or needed.

Each scenario involves a range of types of material and process. Scenario 1 involves the somewhat flexible plant, the gaseous wind, the rigid stake, the 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, but not scenarios 1 and 2, also involves perceptual powers. The facts that the experimenter cannot simply cool a volume of gas that he/she holds in his/her hands or that he/she 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. Because 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 or how to represent the knowledge used in them. I do not believe that these problems will be solved any time soon. 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.

Hayes's 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.

The Naive Physics Manifesto

Commonsense physical reasoning was first and most famously promoted as a domain for AI research by Pat Hayes (1978, p. 2) in the "Naive Physics Manifesto."[2] This paper advocated a research program to develop a formalization of naive physics satisfying the following four criteria: First is thoroughness. "It should cover the whole range of everyday physical phenomena." Second is fidelity. "It should be reasonably detailed." Third is density. "The ratio of facts to concepts should be fairly high." A dense formalization is necessary "to capture the richness of conceptual linking." "Formalizations that are not dense in this way are unsatisfactory since they do not pin down exactly enough the meanings of the tokens they contain." Fourth is 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 has 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 than it might have been" (Hayes 1978, p. 3). Hayes (1978, p. 4) further remarks, "The decision to postpone details of implementation can 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: 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 consists of a preliminary analysis of these clusters. The companion paper, "Ontology for Liquids" (Hayes 1985a), is an in-depth analysis of the liquids cluster.

Finding the proper organization into clusters is considered a key issue 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 gotten it wrong is that it seems hard to say anything useful about the concepts one has proposed, but this can also be the result of having chosen one's concepts badly, having a 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" (Hayes 1978, p. 7).

(I must confess that I personally have never attained the state of grace described in the last sentence. 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 program be carried out by a committee. Each member of the committee would be assigned a particular cluster to formalize. The committee would meet from time to time to integrate their various efforts into a larger theory. This integration would no doubt require that formalizations of clusters be reworked, new clusters be investigated, and old ideas for clusters that prove to be useless be discarded.

One issue that Hayes discusses 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" (Hayes 1978, p. 16), but he does not indicate what the other reasons might be. To my mind, the chief other advantages of naive physics compared to, say, folk psychology or naive social science are, first, the power of real physics, the paradigm of a theory that is comprehensive, exact, and correct. The metatheoretic, mathematical, and logical structures have been studied extensively. 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 parallel with scientific physics gives us an immense body of reliable knowledge on which to draw. Second is that problems of intensionality and self-reference do not arise. Physics is a purely extensional theory. Third is a broad range of practical applications.

Hayes's proposal derives in many key aspects from earlier proposals of John McCarthy's (1968). In particular, the choice of commonsense knowledge as subject matter, the idea of developing knowledge representations independently of implementation, and the choice of first-order logic as a representation language are all taken from McCarthy's previous work. What is chiefly new in Hayes's manifesto is the proposal to restrict the focus to naive physics as opposed to other commonsense domains.

Two Common Misconceptions

There are two common misimpressions of Hayes's 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 readers 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 (for example, Moore [1982]) and Kowalski [1979]) do indeed advocate this view, but Hayes does not, at least not in these papers.[3] He is, in fact, entirely agnostic about how the knowledge should be implemented as data structures or what procedures should manipulate it. Hayes's 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. However, 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. However, 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 (Lifschitz 1986). As 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 theory. (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 should do reasoning in many different directions using whatever partial information it has. Therefore, it is 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. However, these considerations are largely irrelevant to Hayes's 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 focused.

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.

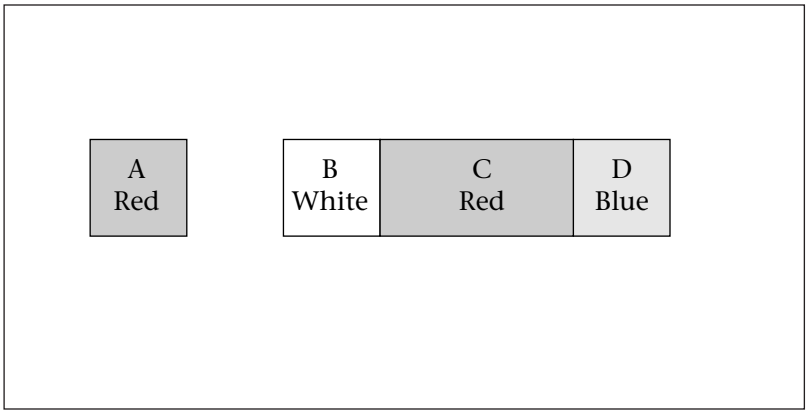

Figure 3. Blocks to Be Represented.

The second common misconception is a little more peculiar. There is a widespread misimpression that if geometric information is represented in first-order logic, then the primitives used must correspond to basic spatial terms in natural language. For example, I have heard it asserted 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 conclude from this supposition 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 supposition. 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 this block, with axes aligned as usual):

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)).

$\forall_X \, block(X) \Rightarrow \exists_Y \, red(Y) \wedge distance(X, Y) < 2.$

In fact, with the exception of probabilistic distributions and fuzzy distributions over space, every representation of spatial or geometric information that I have ever seen can be expressed straightforwardly in a first-order logic over a universe of simple geometric entities. Indeed, in the great majority of representations in use, the ontology can be taken to be Euclidean space, and the language can be restricted to a constraint logic.[4] In particular, all the representations of spatial information that are considered diagrammatic (Glasgow, Narayanan, and Chandrasekaran 1995) can be expressed straightforwardly in first-order logic over Euclidean geometry, and all the inferences considered in Fleck (1996) can be justified in a Euclidean geometry given a suitable statement of the physical axioms. I'm not claiming, of course, that there is necessarily anything to be gained from translating nonlogical representations into logical representations, merely that these alternative representations do not express any kind of information that can't be expressed in first-order logic. There are types of nonspatial information that are impossible or extremely awkward to express in first-order logic, such as uncertain knowledge, metaknowledge, and propositional attitudes, but very, very few declarative representations of spatial information involve any of these problems. (Nondeclarative representations, such as procedural representations, or representations in terms of the state of a neural network do not, of course, translate well into first-order logic.)

Difficulties with the Manifesto

Hayes's 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 work has a quite different flavor from Hayes's proposal; it is algorithmic, rather than declarative, and is increasingly concerned with specialized applications rather than commonsense reasoning (Iwasaki 1997; Weld and de Kleer 1989). 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, and interpreting it narrowly, one could certainly argue that the manifesto and the ontology are the only two papers ever written that fit into Hayes's program.[5]

No doubt the main reason for this neglect is simply that life is short, the project is large, and researchers have had other things to do that seemed more pressing. However, Hayes's project also has fundamental difficulties, and researchers who try to follow Hayes soon face serious obstacles.

It is not 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 (Hayes 1978, p. 29).

> I have deliberately not distinguished between mass and volume. I believe the distinction to be fairly sophisticated (Hayes 1985b, p. 76).

> 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") (Hayes 1978, p. 26).

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 here at least implicitly acknowledges that an individual's beliefs might be inconsistent. (If Newton's first law can be derived by Socratic argument and Gedanken experiments from memories of everyday experience but is also explicitly denied in naive physics, then the closure of the individual's beliefs under "reasonable argument" is inconsistent.)

Trying to define an absolute naive physics raises many difficulties. First, 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" might mean that *A* will spontaneously assert ϕ, *A* will immediately assent to ϕ, *A* will assent to ϕ after Socratic interrogation, *A* will assent to statements that logically entail ϕ, 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 *A*'s actions are more sensible given that ϕ is true than given that it is false. Which is intended here?

Second, the problem of defining *belief* 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.

Third, physical reasoning depends critically on spatial knowledge and spatial reasoning that is difficult or impossible to express in ordinary language. For example, we all know how a screw is shaped, and we all have some understanding of the relation between the shape of a screw and its functions. (This understanding is most easily demonstrated through the methods of considering variants. For example, it is easy to see that a small pit in the surface of the screw will probably have little effect on its behavior, whereas a small bump is likely to be much more troublesome.) However, it is not easy to describe verbally the shape of a screw or explain verbally the connection between its shape and its behavior without using a technical vocabulary unintelligible to most naive subjects. This centrality of spatial knowledge is probably the chief disadvantage of physics, as opposed to other commonsense domains, as a test bed for studying commonsense reasoning.

Fourth, naive physics probably varies substantially from person to person (although Hayes might well be right that it differs less than other branches of commonsense knowledge). Because of the vagueness in defining *naive belief,* it is difficult to be very precise about this variation. However, one can certainly see it in cross-cultural comparisons. For example, many people in various times and places have attributed intentions and mental states to inanimate objects. In modern Western culture, this view is not part of even a naive system of beliefs.

One can try to work around this difficulty by observing that people's beliefs are at least close enough to enable them to communicate and by defining the naive physics we are looking for as the beliefs that are common knowledge within the community. For example, a subject who believes that one sees an object using reflected light and another subject who believes that one sees an object using emanations from the eyes will nonetheless agree that one cannot see through an opaque object. Therefore, if the community contains large numbers of believers in both theories, the naive physics would contain the belief that one cannot see through an opaque object but

Hayes's manifesto was much admired and widely discussed, but it was hardly followed.

would exclude both the theory of reflected light and the theory of ocular emanations as speculative theory. Is it possible to develop a naive physics rich enough to support commonsense inferences on the basis of this kind of common knowledge? The question is important but difficult. Certainly, the central role played by inarticulable spatial knowledge makes this problem more difficult.

Finally, it is not clear that an individual's beliefs are consistent. It depends in part, of course, on how *belief* is defined. An inconsistent belief set cannot be expressed in a single theory in any standard logic (or indeed in most nonmonotonic 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*? Does the distinction between heat and temperature exist? How is one to judge? Pat Hayes (personal communication, 1997) 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.

A particularly difficult issue to judge is the appropriate level of generality. Consider the rule in the cookie-baking domain, "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 following considerations: (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 disjointedly within a region R tends to increase with the area of R. (Note that this is a plausible inference rather than a sound rule.) (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.

One can alternately use rules at an intermediate level of generality (for example, replace the second consideration with 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"). 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 such as *volume, average,* and *cross section,* which are used in our second set of rules, 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," might be one that a child learns first, before any more general formulation, and it might be a rule of thumb that someone baking cookies regularly calls on without doing deeper thought. However, 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 tabletop with books, you will do better to lay them flat and not to stack them; a can of paint will cover a small area more thickly than a large area; and at a further remove, the more people there are sharing a pie, the smaller each person's piece is. To express the general rules that underlie these particular examples, 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 this 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 from a naive understanding. I should say, rather, that teaching these concepts in the classroom is, or should be, mostly a matter of putting concepts that are already understood at the commonsense level into a rigorous setting.

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 rather than stating an axiomatic theory.[6] The argument for this change is as follows: As discussed previously, our main goal in formalizing theories is to characterize or justify the actions of reasoning programs rather than to implement them directly as a rule base. However, the relation of justifying a particular inference or characterizing a particular program is a property of a model, not of a specific axiomatization of this 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 the class of true statements and valid inferences in the model. Second, we are interested in defining a formal language, which delimits an *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's 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.[7]

The following are examples of microworlds: In 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. The microworlds of Forbus (1980) and Sandewall (1989) are similar. In *component-based electronics* (de Kleer and Brown 1985), the world consists of resistors, capacitors, inductors, power sources, and so on, 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. In *rigid-object kinematics,* the world consists of solid, rigid objects constrained by the rules that the shape of an object is fixed, it moves continuously, and two objects do not overlap. In *rigid-object dynamics,* the world consists of medium-sized solid, rigid objects, moving in a uniform gravitational field and interacting through normal forces, friction, and impacts above a fixed ground. In the *kinematics of solid objects and a liquid,* the world consists of solid objects and some quantity of a liquid. The solid objects are constrained by the rules that their shape is fixed, they do not overlap, and their motion is continuous. The liquid is constrained by the rule that its volume is constant, it moves continuously, and it does not overlap the solid objects.

Note the difference from clusters. Of Hayes's clusters, only "liquids" (actually liquids and solids) is a microworld.

We can also contrast microworlds with reasoning architectures, such as QP (Forbus 1985) or ENVISION (de Kleer and Brown 1985). QP and ENVISION do not incorporate any particular physical theory. Rather, each such architecture provides a collection of basic ontological sorts, a restricted language in which physical theories of certain types can be stated, and an algorithm for carrying out certain types of inference. For example, the basic sorts in QP include time instants, time intervals, parameters, and processes. The QP language supplies primitive symbols for direct influence and indirect influence, which have a fixed interpretation. The algorithm carries out qualitative envisioning.

Thus, the development of this kind of program is orthogonal to the microworld methodology. The microworld approach focuses 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's project is in the attitude toward beliefs that are commonsensical but false. These beliefs can be divided into three categories: First are 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. Second are the logical consequences of rules in the first category, 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. Third are beliefs that are just plain wrong, without either of the previous justifications, for example, 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 might well include beliefs of the

first category; indeed at some level, it must, unless we plan to base it on relativistic quantum mechanics. These beliefs are justified as a trade-off of accuracy for speed and simplicity. We are therefore also likely to get beliefs of the second category unless we can block them all using qualification conditions, which is unlikely. The question is whether there is any point in including beliefs of the third category. In Hayes's 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, because 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.

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.

Putting all this together, we arrive at a methodology along the following lines (figure 4):[8]

First, select a microworld, a well-defined, fairly small range of physical behaviors. Second, collect a corpus of inferences in the domain that are both physically correct and would broadly be agreed on as commonsensically obvious. Third, develop a formal model of the domain, a language of primitives with semantics defined in the model, and an axiomatization of the model expressed in the language. Fourth, demonstrate that many of the inferences in the second step can be expressed in the language and justified in the model. A formal proof from the axiomatization might be helpful here. Fifth, 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. Sixth, work toward broadening theories and merging multiple theories together.[9]

Two recent projects of a similar flavor should be mentioned. Ken Forbus (1998) presents a characteristically ambitious proposal to construct a library of foundational qualitative domain theories, containing on the order of 10,000 to 100,000 axioms encoded in first-order logic (I have not been able to get detailed accounts of these domains and so have not been able to make comparisons with the issues discussed here). In a different direction, a recent triad of papers (Lifschitz 1998; Morgenstern 1998; Shanahan 1998) attempts a new method for advancing the state of the art in commonsense reasoning. A complex but narrowly defined inference, such as those used in the original three scenarios, is put forward; separate researchers work independently on developing their own theories for justifying this particular inference; and then hopefully the insights gained from this example can be combined and applied to the next example. These three papers address the problem of characterizing cracking an egg into a bowl.

A Sample Microworld: The Kinematics of Cutting Solid Objects

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

Microworld

The microworld is the kinematics of cutting rigid solid objects (CSO). 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 are being cut. Objects are not created or destroyed except at the moment when one object is sliced through.

The process of cutting is modeled 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 5). This model is rich enough to support many forms 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 object *A*," where the identity of *A* survives in a smaller shape and "slicing object *A* into objects *B* and *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 this 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, because these constraints 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.

Ontology

In developing an ontology, we begin by con-

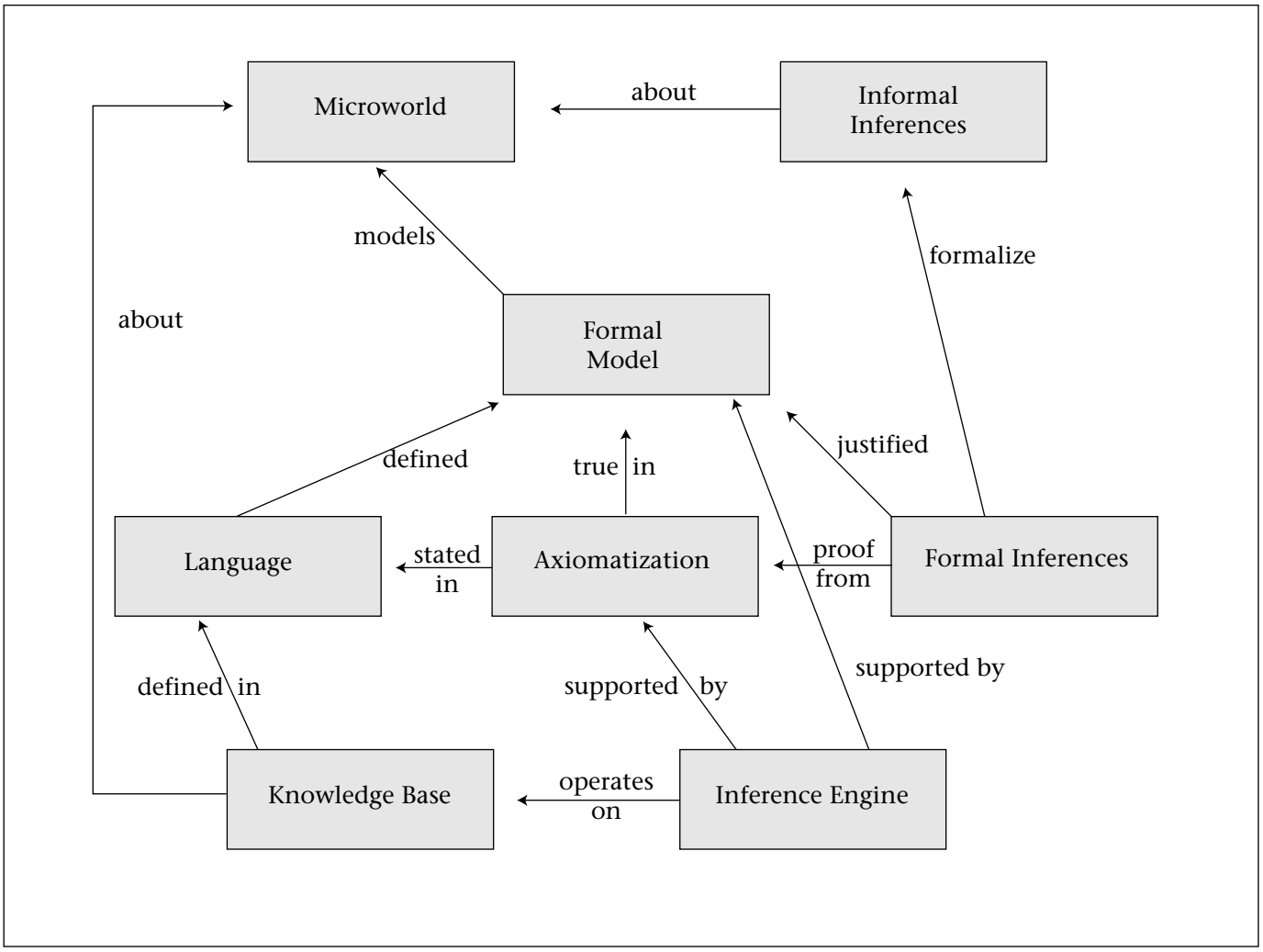

Figure 4. Methodology.

structing a model of time and space. We model time as the real line. A situation is a single instant of time. A *fluent* (McCarthy and Hayes 1969) is an entity whose value changes through time. For example, "the president of the United States" is a fluent whose value in 1791 was George Washington and in 1998 is Bill Clinton. We model space as three-dimensional Euclidean space. Other 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.

We can now formulate two alternative construals of the previous model of cutting. The first, more straightforward approach construes the world in terms of objects, as earlier. 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* becomes disconnected, *O* ceases to be "present" and becomes a "ghost," and two new objects *O*1 and *O*2 cease to be "ghosts" and become "present." Thus, each object can undergo three types of change during its lifetime: First, it is originally created by being sliced off some parent object; second, its shape is gradually modified as it is cut away; third, it is destroyed when its shape is split.

The second construal focuses on chunks of material. A *chunk* is a physically connected piece of material; it is the part of an object that fills some connected, topologically open 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

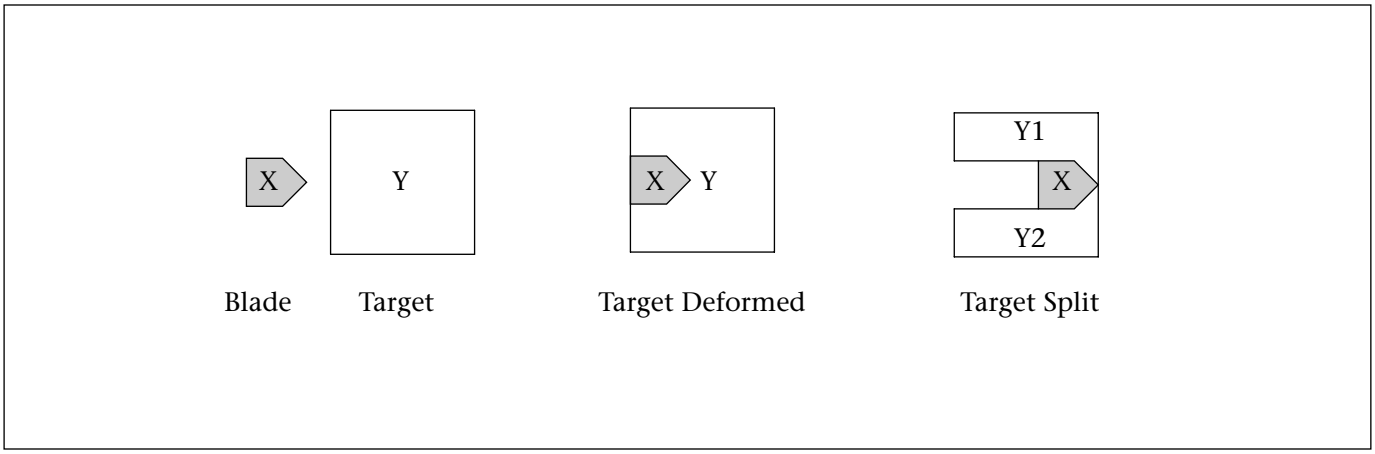

Figure 5. Mutable Object Theory.

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 a blade. Thus, the process of cutting involves the continual destruction of an infinitude of chunks that 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 (that is, one that is either top level or latent) becomes a ghost (figure 6).

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 simplicity can be useful in cases like figure 7. A sculptor is carving away at a pair of stone pieces, of which he/she can see only the nearer parts. In the object theory, this situation is difficult to describe because the sculptor cannot know whether the pieces are, in fact, one object or two; it depends on whether the two pieces are connected, which he/she cannot see. Worse, the two pieces might originally be a single object and then become two when someone splits the connection behind the scene. However, based on the assumption 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.

> Chunk theory and object theory can be proven equivalent (Davis 1993).

We can also 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.* 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.*

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, such as *image,* <, and *continuous,* are defined relative to Euclidean space and real-valued time, as indicated in table 2. 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 example, 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*).

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 computer-aided machinery (CAM) machining programs (Ji and Marefat 1997).

It also supports kinematic inferences of other kinds. For example, Davis (1993, p. 296) gives the proofs of the following statements:

> A blade that starts outside the target cannot carve a purely internal cavity inside the target.

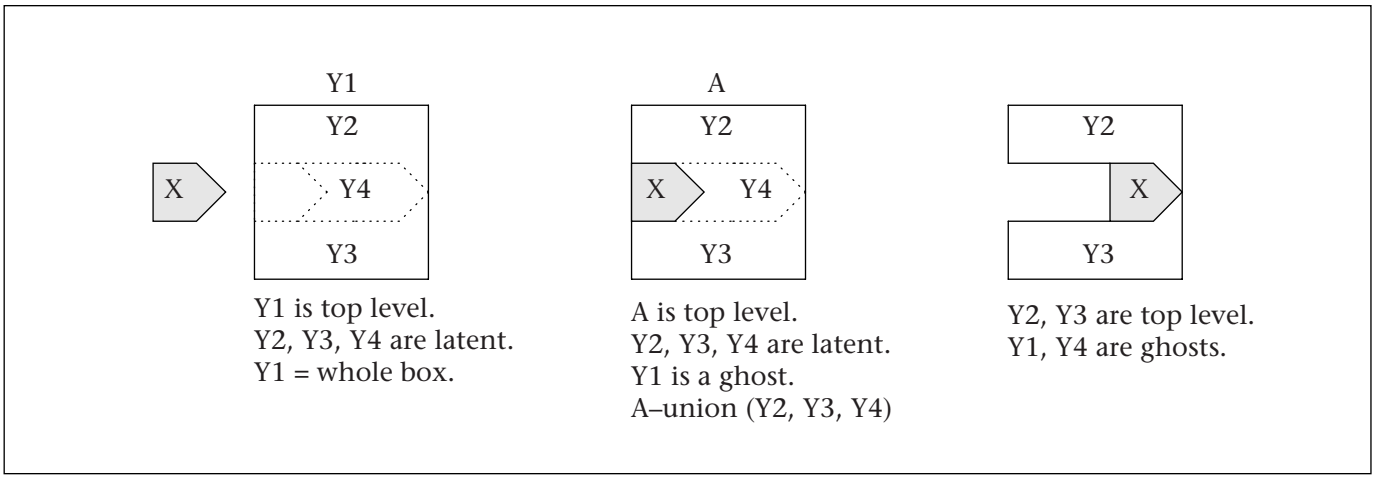

Figure 6. Chunk Theory.

If a convex 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 example, one can conclude that if the dough is cut in the center by a cutter that is a simple, nonclosed 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.

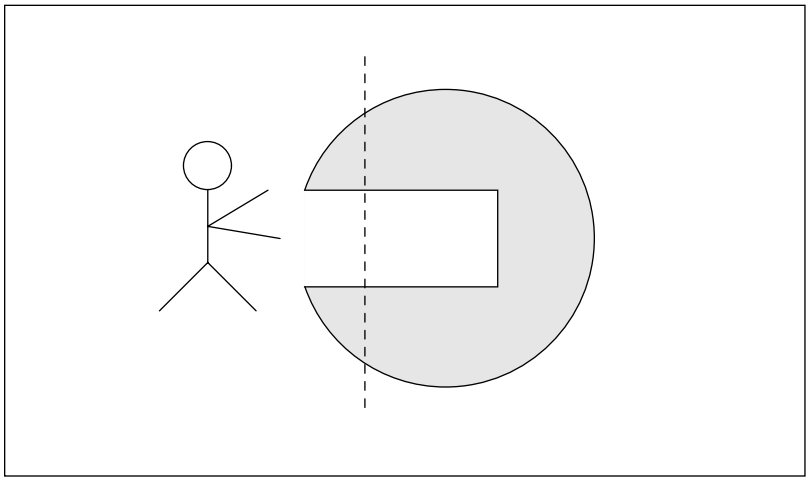

Figure 7. Carving One Object or Two?

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 issues and difficulties.

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 7, or is very much larger than the region being operated on. The theories are certainly rather abstract

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. This is the value of fluent *F* in situation *S*.
*S*1 < *S*2—Predicate. Situation *S*1 precedes *S*2.
just_before(*S, F*)—Predicate. Boolean fluent *F* holds in an open interval ending in *S*.

Spatial

X ∈ *R*—Predicate. Point *X* is in region *R*.
*R*1 ⊂ *R*2—Predicate. Region *R*1 is a proper subset of *R*2.
*R*1 – *R*2—Function. This is the interior of the set difference of *R*1 and *R*2.
intersect(*R*1, *R*2)—Predicate. Region *R*1 intersects *R*2.
∅—Constant. This is the empty region.
good_shape(*R*)—Predicate. Region *R* is nonempty, open, bounded, connected, and equal to the interior of its closure.
image(*M, R*)—Function. This is 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(*R*1, *R*2)—Region *R*1 is a connected component of *R*2.

Physical: Primitive Symbols

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

Physical: Defined Symbols

ghost(*Q*)—Function. This is the fluent of *Q* being a ghost.
place(*Q*)—Function. This is the fluent of the region occupied by *Q* in situation *S*.
blade_swath(*S*1, *S*2, *O*)—Function. This is the swath cut by blades between situations *S*1 and *S*2 relative to the coordinate system attached to object *O*.
destroyed(*S, O*)—Function. Object *O* is destroyed at time *S*.
top_level(*C*)—Function. This is the fluent of chunk *C* being top level.
sub_chunk(*C*1, *C*2)—Predicate. Chunk *C*1 is (nonstrictly) a subchunk of *C*2.

Table 2. Nonlogical Primitives.

and bloodless, mostly, I suspect, because of 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.

Advantages of Microworlds

I next discuss the strengths and flaws of this approach of constructing microworlds to formulate a competence theory. Regrettably, the distinction between strengths and flaws is not always clear-cut. Some apparent strengths can actually be flaws; some apparent flaws can 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 replaces the painfully vague problem, "Is concept–distinction–fact *X* part of naive physics?" with the much more hard-edged, bottom-line, engineering-type question, "Is *X* useful over a given class of inference?" For example, 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 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 ear-

Definitions of Object Theory

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

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

OD.3. $X \in blade\_swath(S1, S2, O) \Leftrightarrow$
$\exists_{S3,OB}\, S1 \le S3 \le S2 \wedge OB \neq O \wedge image(placement(O, S3), X) \in place(OB, S3)$.
(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. $destroyed(S, O) \Leftrightarrow [just\_before(S, material(O)) \wedge \neg\, good\_shape(shape(O, S))]$.
(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. $[material(O1, S) \wedge material(O2, S) \wedge O1 \neq O2] \Rightarrow \neg\, intersect(place(O1, S), place(O2, S))$.
(Two material objects do not overlap.)

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

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

OB.4. $\forall_{S,O}\, shape(O, S) \neq \varnothing \Rightarrow continuous(placement(O), S)$.
(The placement of object O is continuous in any situation S, where the shape of O is nonnull.)

OB.5. $[S1 < S2 \wedge material(O, S1) \wedge just\_before(S2, material(O))] \Rightarrow$
$shape(O, S2) = shape(O, S1) - blade\_swath(S1, S2, O)$.
(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. $[destroyed(S, O) \wedge connected\_component(R, shape(O, S))] \Rightarrow$
$\exists_{OR}\, shape(OR, S) = R \wedge placement(OR, S) = placement(O, S) \wedge$
$just\_before(S, ghost(OR)) \wedge material(OR, S)$.
(If O becomes disconnected or null at S, then each of its connected components becomes material.)

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

OB.8. $[ghost(O, S1) \wedge material(O, S2) \wedge S1 < S2] \Rightarrow$
$\exists_{S3,O3}\, destroyed(S3, O3) \wedge S1 < S3 \le S2 \wedge connected\_component(place(O, S3), place(O3, S3))$.
(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 Object Theory.

lier about whether a painting on the wall is in the room can be avoided. What is actually going on, geometrically and physically, is 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 although Hayes's project requires that naive conclusions be drawn from naive premises, our project requires only that naive conclusions be derivable; the premises need not be formulated in naive terms. Therefore, whereas Hayes's project requires that every concept be examined for its true naiveté 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

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 *subchunk:* $C1$ is a subchunk 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 subchunk 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 subchunk 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 subchunk 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 subchunk 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.

of generality, which we considered previously, is likewise considerably clarified in the new approach. To attain maximal inferential power, one always goes to the highest level of generality that is justified within the scope of the microworld. For example, in the cookie cutter example, one can derive the rule "The thinner you roll the cookie dough, the more cookies you can cut out" from a general theory of volume of regions together with the physical rule that the volume of the dough remains nearly constant while it is 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 formulate this rule at a general level. However, 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 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 the inferences being considered and the precision required. For example, many different theories describe solid objects with varying degrees of accuracy: pure kinematics, quasistatics, Newtonian dynamics of rigid objects, elastic solid objects, and so on. Each of these the-

ories is useful under suitable circumstances. The study of alternative microworlds 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 on getting some particular class of inferences to work without worrying how these inferences will fit with all the rest of the naive physics. Second, it allows much closer ties to practical applications. Most practical AI physical reasoning programs work within a limited scope. For example, many of the programs that do mechanical reasoning (Joskowicz and Sacks 1991; Faltings 1987) work within the microworld of solid-object kinematics or some small extension of it. As I argue further below, this tie to practical applications is valuable for a number of reasons. Third, as the work on automated modeling (Nayak 1994) 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 excess computation. The study of alternative microworlds connects directly to this kind of study.

Focusing on the model, rather than the axiomatization, has the usual advantages of making it much easier to ensure consistency and avoid unintended consequences. As discussed earlier, it is necessary that a concrete extensional model be consistent and precisely defined, thus avoiding much of the conceptual inconsistency and incoherence that can arise in the axiomatic approach.

Dangers and Difficulties of Microworlds

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

The chief problems are these: First, commonsense reasoning is not an autonomous task domain. Second, it is hard to find natural sources for commonsense inferences in a single microworld. Third, the number of potential microworlds is vast, and the methodology provides no guidance for choosing between them. Fourth, the focus on microworlds, rather than axioms, encourages an overemphasis on models that are easily characterized extensionally, on mathematical abstraction and elegance, and on deductive reasoning. Fifth, there is no easy way to extend or integrate microworlds. Sixth, the method involves a great deal of hairsplitting of essentially vacuous issues.

I elaborate on each of these problems individually.

Not a Task Domain

The central objective in the new approach is to develop a competence theory for commonsense physical reasoning. However, 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, and expert systems. 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 rarely improved by any attempt to incorporate commonsense reasoning.

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 final goals in themselves.

Commonsense inference is, thus, an ill-understood module of much larger tasks. It is therefore difficult to be sure what the input and output of this module should be, that is, to decide how a commonsense inference should be formulated to serve the purposes of these larger tasks. In considering commonsense inference for a natural language processor, for example, it is difficult to 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 example, 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, it is unusual to do so, or 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 of a text into

another language often does require this kind of knowledge to choose the proper spatial terms.) In short, the problems of what the representation of a text should be and how world knowledge can be used in linguistic analysis are obscure; therefore, it is difficult to get guidance for commonsense reasoning and representation from linguistic examples (Bloom et al. 1996).

Once we have chosen a microworld, we have to find a collection of inferences within this microworld as a test bed. 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.

Similar ambiguities appear in relating commonsense reasoning to robotics. Here, they take the form of the difficulty in 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 because *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 high-level plan should be the subject of commonsense reasoning or what statements should be inferred about such a plan. Such 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 partial differential equation and a specification of the desired precision and that the output needed is a field of fluid flow of this precision. The difficulty with commonsense reasoning is that there are few instances where we can be really sure what the input and the output should be.

That recent research in knowledge representation suffers from its disconnection from practical applications has been argued by many researchers, including Morgenstern (1997) and Etherington (1997).

No Natural Sources for Single Microworlds

Once we have chosen a microworld, we have to find a collection of inferences within this microworld as a test bed. 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 one assembles 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 this task.[11] However, isolating the commonsense inferences 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 it rarely covers the range of commonsense inference; the inference used tend to be confined to a few types of inference (for example, 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 Three Scenarios, 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 himself/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 might well tend to exaggerate how easily naive subjects can make certain inferences; so, he/she might include, as commonsensical, inferences that are in fact quite difficult.

For example, suppose we want to evaluate how well the model and language for cutting solid objects presented earlier characterize commonsense inference in this domain. 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. However, most of these inferences are of the form, "To create hole *H* in object *O*, move cutter *C* through path *P*." Such inferences can all be justified by a substantially simpler model of cutting, such as one in which each operation with the cutter is taken to be atomic, and a simpler language, such as one in which all geometric descriptions are exact. Most of the other inferences used in the CAM program fall outside the microworld, such as restrictions on the thinness of the parts that can cut out of a given material with a given cutter. We can look at the natural language processing of a technical text describing machining, which 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 the newspaper?

Therefore, 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 a description of motions and outputs final shape descriptions, but such programs have been built by the CAM people much better than I could do. I could set a general-purpose complete theorem prover on the axiom set of tables 3 and 4, together with a set of temporal and geometric axioms, 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 an inference engine that will work efficiently over the space of commonsensically obvious inferences, but I don't know what this 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 great value. However, a program that can do commonsense reasoning about cutting solid objects, or similarly narrow domains, would be of little value. Therefore, it is difficult to know what any of these programs should do about cutting solid objects. We don't know what a 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 occasionally be called on to carry out an inference that is both nontrivial 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, although these considerations are not to be sneezed at. Far more importantly, it means that there is almost no way to guide research in microworlds or 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 nearly complete.

Innumerable Microworlds

Hayes's 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 are endless variations on the blocks world: Blocks can stack in towers one on one, they can be rectangular of varying sizes, or they can have more general shapes. Time and space can be continuous or discrete; there can be one hand or many hands; if many, they can work one at a time or concurrently, and they can interact in any of several ways; and so on. This wealth of microworlds is useful for the teacher giving a class in knowledge representation who needs simple examples to assign, but for the researcher, only a few of these merit any study. The methodology described earlier does not give one any clue about when the analysis of a new microworld is worthwhile. The choice of where to invest energy is left entirely to the judgment of the researcher, and knowledge representation research has always been remarkably apt to leave the great ocean of truth undiscovered but crowd around an empty Clorox bottle on the beach.

Hayes's 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.

Overemphasis on Extensional Models

The model-based methodology strongly focuses attention 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 rule:

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

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 *function*. It is also unlikely to be found as a sample commonsense inference by the method of proposing variants because it is too general.

Excessive Mathematization

… the model-based methodology leads to an excessive interest in constructing elegant and minimal mathematical models rather than expressive, messy models.

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 earlier 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 are complicated.) Moreover, these rules and constraints are not disconnected arbitrary facts but are deeply interconnected. For example, anyone who has observed the processes of butter being sliced and wood being sawed will expect, from the nature of the processes and the materials, that butter can be sliced more thinly than wood can be sawed. However, it is not easy to formulate the general rules that give rise to this expectation.

The researcher who wants to move forward producing models will therefore tend to avoid this kind of microworld because 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 would also be harder to "sell" as legitimate research; they look like a mere translation of random obvious statements into an arbitrary formal notation. In fact, the immense gap between a mere translation of random statements and a coherent theory is no smaller in a complex theory than in a simple one, but the coherence of the complex theory is harder to achieve, convey, and 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 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. These theorems are often of doubtful relevance. A 22-page proof that two theories of cutting are mathematically equivalent (Davis 1993) certainly does not represent any cognitive activity that anyone (except myself) has ever carried out or any computational activity that any program is ever likely to carry out. Now, of course, I can 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 with regard to 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. Previously, we suggested that the special rule, "The thinner you roll the cookie 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 a fixed volume and that cutting out cookies corresponds to dividing the region of the dough into vertical cylinders with some fixed cross section. However, this generalization fails to capture the causal direction of the special 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 conscious effort.

Too Much Stress on Deduction

Being centered around semantic consequence, the microworld 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, but would be difficult to integrate with such theories as default reasoning, reasoning by analogy, and case-based reasoning.

Extending and Combining Microworlds

An advantage of Hayes's project is that because the aim at every step is always a complete theory of naive physics and because every axiom of every cluster is true relative to this overall theory, once you have correctly formulated an axiom or a cluster, you can count on it and keep it. If it is true, it remains true. In the microworld approach, by contrast, a model that has been constructed to characterize a narrow microworld does not usually apply in a broader world. Models, theories, languages, and axioms almost always require some revision in going from a narrower to a wider setting and might well require complete reworking from scratch.

Let us first consider a case where the extension of one model to a richer model has a straightforward logical structure. The kinematic theory of rigid solid objects (KRSO) can be extended to a dynamic theory by adding mass, force, momentum, and so on, and imposing Newton's laws. This extension is what Giunchiglia 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 decreasing* (Nayak and Levy 1995);[12] 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.[13]

A more complex example is the extension of KRSO to the theory of CSO described earlier. This is a model-increasing extension; any history consistent with the KRSO theory is also consistent with CSO. Correspondingly, it is a theorem-decreasing extension. This seems a little odd because the CSO contains all kinds of axiom and inference about cutting that don't apply in KRSO, but actually these are all vacuously true in the KRSO case. For example, it is true in KRSO 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 of feasibility such as "you can cut through an apple with a knife" are not part of CSO as I have defined it.)

However, this simple characterization requires a significant qualification. By starting with CSO, it is easy to construct KRSO as a special case by adding to the mutable objects theory the axiom that the shape of an object is constant or adding to the chunk theory the axiom that all chunks are eternal. Going from KRSO to CSO, which is the more likely order of development, is much more difficult. In the natural logical statement of KRSO, shown in tables 5 to 7, there is no need for the fluent *material*(*O*) because all objects are eternal, and the function *shape*(*O*) maps an object *O* to a spatial region, rather than to a fluent, because the shape of an object is fixed. Thus, developing the mutable object theory of CSO from KRSO requires significant reworkings of the conceptualization, ontology, and language in addition to changing the axioms. (Note that only one of the axioms from table 7 survives unchanged from table 3.) Developing chunk theory requires even greater ontological changes, although, curiously, fewer axiomatic changes (three axioms from table 7 appear in the same form in table 4).

Similar difficulties are encountered in trying to combine two microworlds; all too often, one finds that each microworld depends on assumptions that are violated in the other. Let me discuss an example that has been fretting me for some years. I have a theory of CSO. I also have a theory of strings, presented briefly in Davis (1995). The form of this theory is determined by the following considerations:

First, the length of a string is constant. Second, strings are flexible. Third, making strings one-dimensional curves is problematic, though tempting. For example, 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 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, although thin. Fourth, the diameter of a string is generally much less than

Being centered around semantic consequence, the microworld approach tends to focus exclusively on deductive reasoning.

Sort	Letter
Spatial regions (set of points)	*R*
Rigid mappings	*M*
Temporal situations	*S*
Fluents	*F*
Objects	*O*

Table 5. Logical Sorts in the Kinematic Theory of Rigid Solid Objects (KRSO).

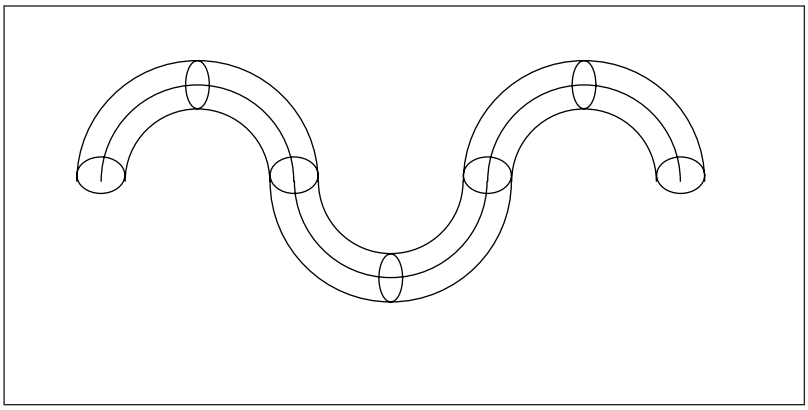

Figure 8. Theory of String.
The central curve is the core of the string, and the small circles are cross-sections.

its length, and the shape of its cross section is unimportant for most purposes. Fifth, we want 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 much the same from one string to another.

To accommodate these constraints, I propose the following kinematic theory of strings and solid objects (figure 8):

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 cross section of the string perpendicular to the core is a circle of radius R; that is, the extension of the string occupies all points of the form $q + \nabla\hat{N}$, where q is a point in the core C; N is normal to the curve C at q and $\Delta \leq R$. The string observes the following constraints:

First, the string moves continuously. Second, the string does not overlap any solid rigid object. Third, the string does not overlap any other string. Fourth, the string does not overlap itself. That is, there cannot be two distinct points q_1 and q_2 on curve C; two normals $\hat{N}_1$ and $\hat{N}_2$ to C at q_1 and q_2; and two quantities Δ_1 and $\Delta_1 < R$, such that

$q_1+\Delta_1\hat{N}^1= q_2 + \Delta_2\hat{N}_2$

This theory is reasonably straightforward and integrates directly with KRSO. 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 they are 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 earlier, assumes that a string has a circular cross section everywhere.

Now, there might 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 stringlike 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 varies depending on what the string is made of. Hence, it is not surprising that modeling cutting string is not a simple extension of modeling string.

However, cutting string is not, after all, an 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. Three related reactions to this 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 cut through halfway; when we start to cut a string, we usually complete the job. The third is that the requirement that strings with a circular cross section should be dropped; there are many strings, including sneaker laces and strings that have been partially cut through, that do not.

The first of these reactions is actually a fallacy, based on the ease with which human reasoners solve and, therefore, ignore the frame

Temporal

value_in(*S, F*)—Function. This is the value of fluent *F* in situation *S*.

Spatial

intersect(*R*1, *R*2)—Predicate. Region *R*1 intersects *R*2.

good_shape(*R*)—Predicate. Region *R* is nonempty, open, bounded, connected, and equal to the interior of its closure.

image(*M, R*)—Function. This is 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.

Physical: Primitive Symbols

placement(*Q*)—Function. This is the fluent of the mapping from the shape of *Q* to the place of *Q*.

shape(*O*)—*Function.* This is the point set occupied by *O* in a standard orientation.

Physical: Defined Symbols

place(*Q*)—Function. This is the fluent of the region occupied by *Q* in situation *S*.

Table 6. Nonlogical Primitives in the Kinematic Theory of Rigid Solid Objects (KRSO).

K.1. $place(O, S) = image(placement(O, S), shape(O))$.
Definition of *place:* The region occupied by *O* in *S* is the image of its shape under its placement.)

K.2. $O1 \neq O2 \Rightarrow \neg\, intersect(place(O1, S), place(O2, S))$.
(Two objects do not overlap.)

K.3. $good\_shape(shape(O))$.
(Every object has a good shape.)

K.4. $continuous(placement(O), S)$.
(The placement of object *O* is continuous in any situation *S*.)

Table 7. Axioms of the Kinematic Theory of Rigid Solid Objects (KRSO).

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. Thus, the reaction that we don't care is presupposing some strong constraints on the behavior of the string while it is 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. Such a model can be developed 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 might 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 pen-

etrated, then what we get is precisely the previous model, that the string is split as soon as the blade enters it. (Chunk theory also allows a more elegant expression of the rule 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 this sense, it is a reasonable competence theory.[14] Unlike the microworlds we 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 (that is, none). Then, this model will do the right thing as long as we never have to reason about incomplete cuts or the state of the string during cutting. The fact that it is easier to construct such an overly specified model, rather than just characterize correctly the starting and ending states and the interaction with the rest of the world, illustrates another defect of the model-based methodology: It tends to generate overly specific theories.

The third suggestion, that we should allow strings with noncircular cross sections, has in its favor that it is true, and it will have to be accommodated in an ultimate commonsense theory. However, the theory of noncylindrical strings is significantly more complicated than the theory of cylindrical strings for a number of reasons: First, noncylindrical strings can twist; in cylindrical strings, twist is invisible and can therefore be ignored. Second, noncylindrical strings are more restricted in the shapes they can attain. For example, it is not possible to wind a sneaker lace tightly in the plane of the lace itself without buckling because the outer diameter of the lace becomes so much longer than the inner diameter. The microworld approach has value insofar as it allows us to focus on a natural class of issues, and it would seem natural that we should be able to reason about the common and familiar process of cutting string without getting involved in all the rare and specialized issues of oddly shaped string.

Hairsplitting

By this stage of the article, few readers will need more illustrations of this point! A little earlier, we were patting ourselves on the back because we could avoid two-hour discussions on the meaning of *in*; although this 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 case and anomaly be resolved.

It is somewhat tolerable that AI should have trouble with real borderline cases such as whether a platypus is a mammal, whether glass is a solid, or what is the nature of an impulse. After all, scientists and engineers who study this kind of issue also work hard to resolve such 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 noninsistence 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 his/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 such as cutting string, they are trying to deal with truth, justice, and beauty. However, 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.

The Role of Microworlds in the Larger Scheme of Things

When we are all done—when we have encoded all commonsense physical knowledge in a declarative knowledge base and implemented all commonsense physical reasoning in an

inference engine over the knowledge base—how will our work on microworlds be reflected in the final product? Three possibilities come to mind:

One possibility is that we will attain Hayes's dream of a single consistent theory that incorporates all commonsense physical knowledge and supports all commonsense physical inferences. In this case, our microworlds would certainly serve no intrinsic logical function. They would survive at most as organizational structures, clusters within the knowledge base supporting efficient retrieval. More likely, in view of our earlier observations of their characteristic inextensibility, they would simply vanish. Their whole function in the research project, then, would have been as stepping stones and training exercises for the eventual theory. Certainly, there would be no point in ever going backwards; once a more comprehensive microworld had satisfactorily been formulated, there would be no interest in considering special cases.

A second possibility, along the lines of Addanki, Cremonini, and Penberthy (1989), is that the final knowledge base would be structured entirely in terms of mutually inconsistent microworlds, with no overall theory. There would be rules at the metalevel for choosing the microworld suitable to a given problem or resolving conflicts when different microworlds gave different answers, but there would be no object-level theory that would serve as the final court of appeals in such cases.

A third possibility combines the two possibilities. There is a structure of microworlds, integrated through metarules, but at the top of this structure is a single Hayesian theory to which all questions can ultimately be referred (figure 9). The microworlds approximate the overall theory and are computationally more tractable. This possibility can be viewed as a special case of the second structure in which there happens to be a single overarching top-level microworld. Alternatively, it can be viewed as an instance of the first structure by taking the overall theory to be logically primary and viewing calculations involving the microworlds as approximation heuristics to the overall theory.

Undoubtedly, we currently know much too little to predict which, if any, of these three possibilities will win out. However, a few pros and cons can be observed.

The first structure, of a single comprehensive theory, is certainly the simplest from a logical standpoint. Indeed, the idea of using multiple worlds runs seriously counter to goals of a declarative representation or a knowledge-based analysis. This conflict becomes particularly evident in cases where you want to use two conflicting models in a single problem, either to describe two different objects in the problem, two different times, two different places, two different scales of granularity, or two different interactions. For example, to calculate the tides, you first calculate the motions of the earth and the moon around the sun as if they were point objects, then treat the earth as a solid of complex shape with bodies of water. The reasoner must somehow keep the inferences from each microworld within the range of its applicability and avoid making inferences from these microworlds that make nonsense of the problem being solved; this idea of a limited range of inference is not one that works easily within the standard view of a knowledge-based system.

This difficulty is not necessarily alleviated by formally incorporating a microworld structure inside a single first-order theory, as is done in the theory of contexts (McCarthy 1993). The fundamental problem of allowing different theories to interact in useful ways but not in destructive ways remains difficult, however the theories are combined.

Moreover, I have not found any convincing arguments that a structure of alternative microworlds is a particularly plausible cognitive model of commonsense physical reasoning. I do not know of any cases where commonsense reasoning seems to require the combination of two conflicting models. I suspect that in developing a commonsense physical reasoner, our ultimate aim should be something like Hayes's uniform, comprehensive theory. Therefore, I should tend to stress the steady expansion of the scope and detail of our theories rather than pursue such virtues as simplicity or tractability.

By contrast, in developing an automated reasoner for expert scientific or engineering reasoning, the idea of using alternative microworlds approximating a single ultimate correct theory seems much more promising. In formulating and solving a problem, a scientist-engineer will almost always simplify, abstract, and approximate; he/she can generally describe the approximations he/she is making and, to some extent, explain why they will simplify the problem and why he/she expects that the answer will still be useful. A large part of scientific and engineering training has to do with learning a library of useful approximations and abstractions and learning how to apply these to different problems. Indeed, there is an active research area trying to develop an account of the relations between

... in developing an automated reasoner for expert scientific or engineering reasoning, the idea of using alternative microworlds approximating a single ultimate correct theory seems much more promising.

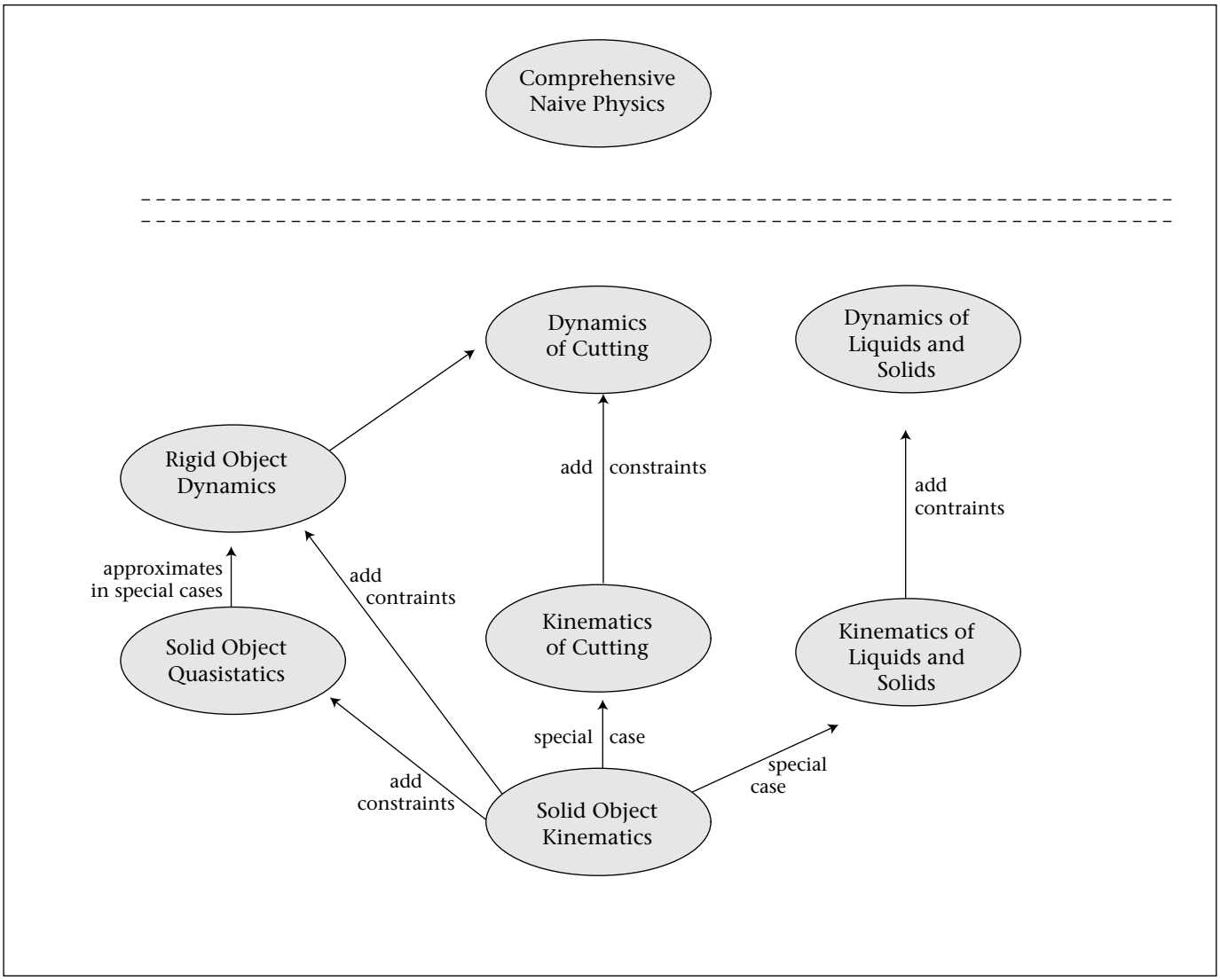

Figure 9. Part of a Structure of Microworlds.

microworlds and see how a structure of microworlds can be used in automated expert physical reasoning.

I suspect, however, that characterizing the formal relations between microworlds is not central to the use of approximation and abstraction in physical reasoning. Approximation and abstraction in physical reasoning takes many forms: Objects of complex characteristics can be approximated by objects of simpler characteristics (for example, a real resistor by an ideal resistor or a curved object by a polyhedral object), a group of objects can be approximated by a single object (figure 10), and a group of simple objects can be approximated by a single complex object (for example, a chain of rigid links by a string). The case where a complex microworld is approximated by a simpler microworld is just one of many cases, and it is not obvious that it is a particularly important special case. Also, it is a mistake to assume, as is sometimes done (for example, Weld [1992]), that a problem is easier to solve in a formally simpler theory than in a more complex theory; in many important cases, the reverse holds. For example, solid-object dynamics is a formally simpler theory if friction is excluded than if it is included, but in many problems, such as the system in figure 11, prediction is easy in a theory with friction—the system remains static—but difficult in a frictionless theory.

Because many problems in physical reasoning can be solved within the scope of a single microworld, the development of microworlds remains useful in developing automated expert physical reasoners. I suspect, however, that the study of the relationships between

microworlds and the manipulation of microworld assumptions will be much less important in developing sophisticated reasoning techniques.

Conclusions

Despite all these difficulties and objections and despite the increasing impatience of the AI community with laboriously hand-coded knowledge-based systems (for example, Charniak [1993]), I find our original scenarios—the staked plant, the cookie dough, the baby bottles, and a myriad of 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 we discussed, 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, reasoning will rely more on plausible inference, the spatial component of reasoning will be both more complex and less clearly defined, and immediate connections to useful applications will become fewer. However, if we have patience enough to stick with it, we should eventually have a remarkable theory.

Acknowledgments

This article originated in invited lectures given at the University of York and the University of

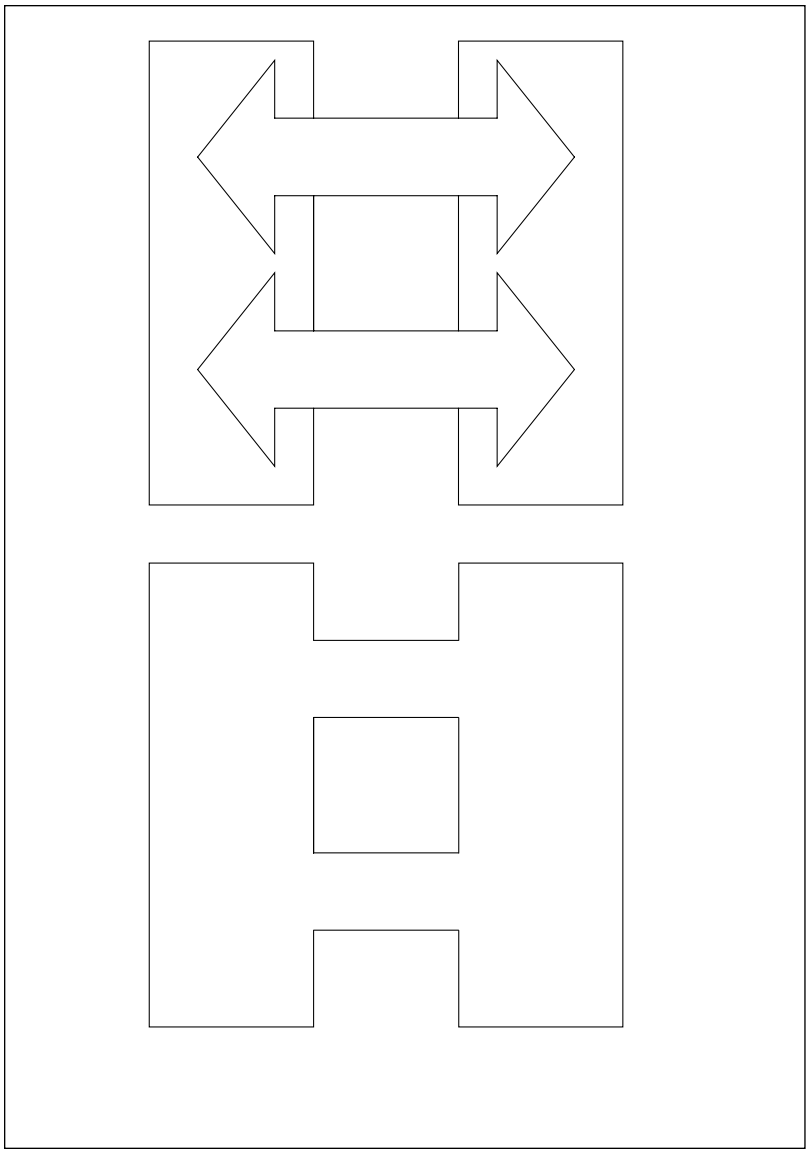

Figure 10. Abstraction of Structure: Several Objects Become a Single Object.

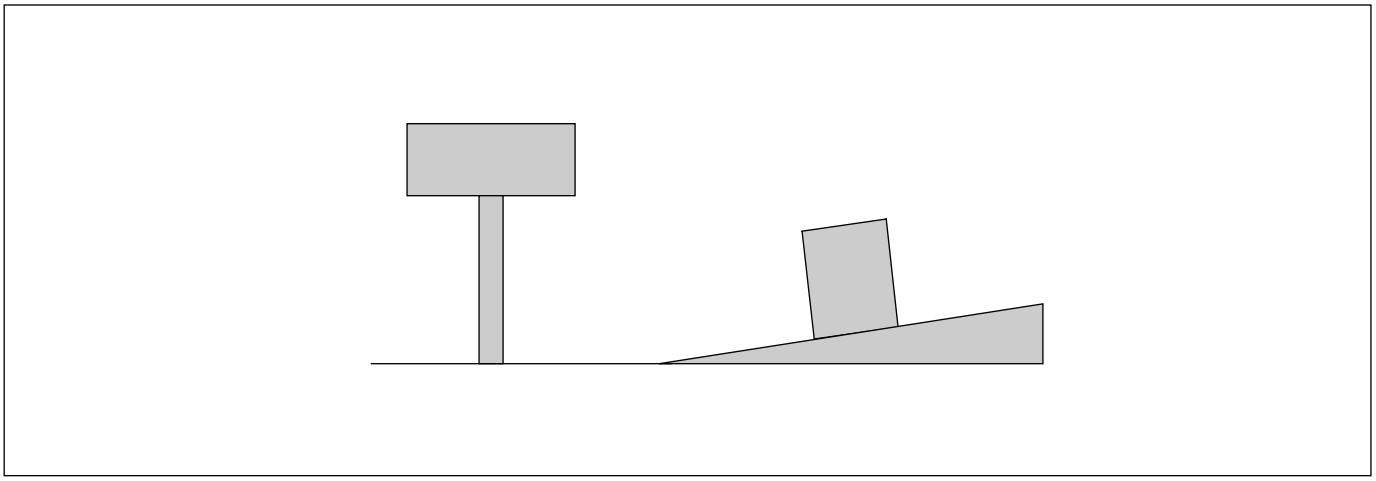

Figure 11. A Simple System with Friction

Leeds. Thanks to my hosts there, Tony Cohn and Alan Frisch. Thanks also to Philip Davis, Leora Morgenstern, Pat Hayes, and Drew McDermott for encouragement and helpful suggestions. In particular, scenario 2 was suggested by Leora Morgenstern. The intellectual debt of this article to Pat Hayes should be entirely evident. The writing of this article was supported by National Science Foundation grants IRI-9300446 and IRI-9625859.

Notes

1. I tried this experiment three times. Twice, the entire baby bottle collapsed under 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.

2. All quotations in this section are taken from Hayes (1978). The published version of this is always cited as Hayes (1979); however, I have never actually set eyes on this version, and I don't know what changes might have been made before publication. The later version (Hayes 1985b) is a substantially different paper.

3. Even in the paper "In Defense of Logic" (Hayes 1977), 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.

4. A constraint logic is the conjunction of atomic ground sentences, without negation, disjunction, or quantification.

5. Schmolze (1986) should also be mentioned.

6. I generally use *model* in this article in the sense common in physical reasoning research: A model is an abstract structure that mirrors some of the significant properties of a physical microworld. The notion of "conceptualization" in Guarino (1998) is similar. This meaning is somewhat different from the meaning of the term in metalogic. When I need the term from metalogic, I say so specifically.

7. The term *microworlds* goes back in AI research at least as far as the early 1970s (Minsky and Papert 1970). These microworlds, however, had quite a different purpose; they were simplified test beds for exploring such issues as inference, search, planning, and learning. CYC in its later versions (Lenat and Guha 1993) is the most notable recent exemplar of the use of microtheories. CYC microtheories are axiom based rather than model based.

8. 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 argue from a shift from an axiomatic to a model-based analysis in automated reasoning.

9. 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.

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 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.

11. Interestingly, the original plan for CYC (Lenat, Prakash, and Shepherd 1986) was to express the background knowledge needed to understand encyclopedia articles (hence, the name); they later report "that use of external written materials has become increasingly rare" (Lenat and Guha 1993, p. 152).

12. This is *model* in the strict metalogical sense.

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

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

References

Addanki, S.; Cremonini, R.; and Penberthy, J. S. 1989. Reasoning about Assumptions in Graphs of Models. In *Readings in Qualitative Reasoning about Physical Systems,* eds. D. Weld and J. de Kleer, 546–562. San Francisco, Calif.: Morgan Kaufmann.

Bloom, P.; Peterson, M. A.; Nadel, L.; and Garrett, M. F., eds. 1996. *Language and Space*. Cambridge, Mass.: MIT Press.

Charniak, E. 1993. *Statistical Language Learning*. Cambridge, Mass.: MIT Press.

Charniak, E., and McDermott, D. 1985. *Introduction to Artificial Intelligence*. Reading, Mass: Addison Wesley.

Davis, E. 1995. Approximation and Abstraction in Solid-Object Kinematics. Technical Report, 706, Computer Science Department, New York University.

Davis, E. 1993. The Kinematics of Cutting Solid Objects. *Annals of Mathematics and Artificial Intelligence* 9(3–4): 253–305.

Davis, E. 1990. *Representations of Commonsense Knowledge*. San Francisco, Calif.: Morgan Kaufmann.

de Kleer, J. 1977. Multiple Representations of Knowledge in a Mechanics Problem Solver. In Proceedings of the Fifth International Joint Conference on Artificial Intelligence, 299–304. Menlo Park, Calif: International Joint Conferences on Artificial Intelligence.

de Kleer, J., and Brown, J. S. 1985. A Qualitative Physics Based on Confluences. In *Qualitative Reasoning about Physical Systems,* ed. D. Bobrow, 7–84. Cambridge, Mass.: MIT Press.

Dreyfus, H. 1979. *What Computers Can't Do: The Limits of Artificial Intelligence*. New York: Harper and Row.

Etherington, D. 1997. What Does Knowledge Representation Have to Say to Artificial Intelligence? In Proceedings of the Fourteenth National Conference on Artificial Intelligence, 762. Menlo Park, Calif: American Association for Artificial Intelligence.

Faltings, B. 1987. Qualitative Kinematics in Mechanisms. In Proceedings of the Tenth International Joint Conference on Artificial Intelligence, 436–442. Menlo Park, Calif: International Joint Conferences on Artificial Intelligence.

Fleck, M. 1996. The Topology of Boundaries. *Artificial Intelligence* 80(1): 1–27.

Forbus, K. 1998. Building and Using Large Common Sense Knowledge Bases. Available at www.qrg.ils.nwu.edu/projects/HPKB/.

Forbus, K. 1985. Qualitative Process Theory. In *Qualitative Reasoning about Physical Systems,* ed. D. Bobrow, 85–168. Cambridge, Mass.: MIT Press.

Forbus, K. 1980. Spatial and Qualitative Aspects of Reasoning about Motion. In Proceedings of the First National Conference on Artificial Intelligence. Menlo Park, Calif:

American Association for Artificial Intelligence.

Giunchiglia, F., and Walsh, T. 1992. A Theory of Abstraction. *Artificial Intelligence* 57:323–389.

Glasgow, J.; Narayanan, N. H.; and Chandrasekaran, B. 1995. *Diagrammatic Reasoning.* Cambridge, Mass.: MIT Press.

Guarino, N. 1998. Formal Ontology and Information Systems. In *Formal Ontology in Information Systems,* ed. N. Guarino. Amsterdam, The Netherlands: IOS.

Halpern, J., and Vardi, M. 1991. Model Checking versus Theorem Proving: A Manifesto. In Proceedings of the Second International Conference on Knowledge Representation and Reasoning, 325–334. San Francisco, Calif: Morgan Kaufmann.

Hayes, P. 1985a. Naive Physics 1: Ontology for Liquids. In *Formal Theories of the Commonsense World,* eds. J. Hobbs and R. Moore, 71–108. Norwood, N.J.: Ablex.

Hayes, P. 1985b. The Second Naive Physics Manifesto. In *Formal Theories of the Commonsense World,* ed. J. Hobbs and R. Moore, 1–36. Norwood, N.J.: Ablex.

Hayes, P. 1979. The Naive Physics Manifesto. In *Expert Systems in the Microelectronic Age,* ed. D. Michie. Edinburgh: Edinburgh University Press.

Hayes, P. 1978. The Naive Physics Manifesto. Working paper, Department of Computer Science, University of Essex.

Hayes, P. 1977. In Defense of Logic. In Proceedings of the Fifth International Joint Conference on Artificial Intelligence, 559–565. Menlo Park, Calif: International Joint Conferences on Artificial Intelligence.

Iwasaki, Y., ed. 1997. Qualitative Reasoning: When Precise Data Isn't What You Need. *IEEE Expert* (Special Issue on Intelligent Systems and Their Applications) 12(3).

Ji, Q., and Marefat, M. 1997. Machine Interpretation of CAD Data for Manufacturing Applications. *ACM Computing Surveys* 29(3): 264–311.

Joskowicz, L., and Sacks, E. 1991. Computational Kinematics. *Artificial Intelligence* 51:381–416.

Kowalski, R. 1979. Algorithm = Logic + Control. *Communications of the ACM* 22:424–436.

Lenat, D., and Guha, R. V. 1993. RE:CYCLING Paper Reviews. *Artificial Intelligence* 61:149–174.

Lenat, D.; Prakash, M.; and Shepherd, M. 1986. CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge-Acquisition Bottlenecks. *AI Magazine* 6:65–85.

Lifschitz, V. 1998. Cracking an Egg: An Exercise in Formalizing Commonsense Reasoning. Paper presented at the Fourth Symposium on Logical Formalizations of Commonsense Reasoning, 7–9 January, London, U.K.

Lifschitz, V. 1986. On the Semantics of STRIPS. In *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop, Timberline, Oregon,* 1–9. San Francisco, Calif: Morgan Kaufmann.

McCarthy, J. 1993. Notes on Formalizing Context. In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, 559–565. Menlo Park, Calif: International Joint Conferences on Artificial Intelligence.

McCarthy, J. 1968. Programs with Common Sense. In *Semantic Information Processing,* ed. M. Minsky, 403–418. Cambridge, Mass.: MIT Press.

McCarthy, J., and Hayes, P. 1969. Some Philosophical Problems from the Standpoint of Artificial Intelligence. In *Machine Intelligence 4,* eds. B. Metzler and D. Michie, 463–502. Edinburgh: Edinburgh University Press.

McCloskey, M. 1983. Naive Theories of Motion. In *Mental Models,* eds. D. Gentner and A. Stevens, 299–324. Hillsdale, N.J.: Lawrence Erlbaum.

McDermott, D. 1987. A Critique of Pure Reason. *Computational Intelligence* 3:151–160.

McDermott, D. 1978. Tarskian Semantics, or No Notation without Denotation! *Cognitive Science* 2:277–282.

Miller, R., and Morgenstern, L. 1998. Commonsense Problem Page. Available at www-formal.stanford.edu/leora/cs/.

Morgenstern, L. 1998. Formalizing a Problem in Commonsense Physical Reasoning: The Egg-Cracking Problem. Paper presented at the Fourth Symposium on Logical Formalizations of Commonsense Reasoning, 7–9 January, London, U.K.

Morgenstern, L. 1997. Inheritance Comes of Age: Applying Nonmonotonic Techniques to Problems in Industry. In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, 299–304. Menlo Park, Calif: International Joint Conferences on Artificial Intelligence.

Moore, R. 1982. The Role of Logic in Knowledge Representations and Commonsense Reasoning. In Proceedings of the Second National Conference on Artificial Intelligence, 428–433. Menlo Park, Calif: American Association for Artificial Intelligence.

Nayak, P. 1994. Causal Approximations. *Artificial Intelligence* 70:277–334.

Nayak, P., and Levy, A. 1995. A Semantic Theory of Abstractions. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 196–202. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.

Newell, A. 1980. The Knowledge Level. *AI Magazine* 2(2): 1–20.

Sandewall, E. 1989. Combining Logic and Differential Equations for Describing Real-World Systems. In *Proceedings of the First International Conference on Knowledge Representation and Reasoning,* 412–420. San Francisco, Calif: Morgan Kaufmann.

Schmolze, J. 1986. Physics for Robots. In Proceedings of the Fifth National Conference on Artificial Intelligence, 44–50. Menlo Park, Calif: American Association for Artificial Intelligence.

Shakhashiri, B. Z. 1985. *Chemical Demonstrations: A Handbook for Teachers of Chemistry.* Madison, Wis.: University of Wisconsin Press.

Shanahan, M. 1998. A Logical Formalisation of Ernie Davis's Egg-Cracking Problem. Paper presented at the Fourth Symposium on Logical Formalizations of Commonsense Reasoning, 7–9 January, London, U.K.

Tuttle, M. S. 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* 61:121–148.

Weld, D. 1992. Reasoning about Model Accuracy. *Artificial Intelligence* 56:255–300.

Weld, D., and de Kleer, J. 1989. *Readings in Qualitative Reasoning about Physical Systems.* San Francisco, Calif: Morgan Kaufmann.

Ernest Davis received his B.Sc. in mathematics from the Massachusetts Institute of Technology and his Ph.D. in computer science from Yale University, under the advisement of Drew McDermott. He is currently associate professor of computer science at New York University. His research area is in representation of commonsense knowledge and commonsense reasoning, particularly physical and spatial reasoning. He is the author of the textbook *Representations of Commonsense Knowledge* (San Francisco, Calif.: Morgan Kaufmann, 1990). His e-mail address is davise@cs.nyu.edu.