COGNITIVE SCIENCE TECHNICAL REPORT

University of Rochester
Rochester, NY 14627

URCS - 10
October 1983

THE SECOND NAIVE PHYSICS MANIFESTO

by

Patrick J. Hayes
Cognitive Science
University of Rochester

THE SECOND NAIVE PHYSICS MANIFESTO

Patrick J. Hayes

Ø. PREFACE

Five years ago I wrote a paper, "The Naive Physics Manifesto", complaining about AI's emphasis on toy worlds and urging the field to put away childish things by building large-scale formalisations, suggesting in particular that a suitable initial project would be a formalisation of our knowledge of the everyday physical world: of naive physics (NP). At that time, I felt rather alone in making such a suggestion (which is why the paper had such a proselytizing tone) and quite optimistic that success in even this ambitious a project could be achieved in a reasonable time scale. As this volume testifies, both feelings are no longer appropriate. There is a lot of work going on, and there is more to be done than I had forseen. A whole layer of professionalism has emerged, for example, in the business of finding out just what people's intuitive ideas are about such matters as falling rocks or evaporating liquids, a matter I had relegated to disciplined introspection. In 1978, I predicted that the overall task was an order of magnitude (but not ten orders of magnitude) more difficult than any that had been undertaken so far. I now think that two or three orders of magnitude is a better estimate. It's still not impossible, though.

My old paper now seems dated and, in places, inappropriately naive on some deep issues. The following is a revised version which attempts to correct some of these shortcomings,and repeats the points which need repeating because

nobody seems to have taken any notice of them.

If, dear reader, you have seen the original, please understand that this is a revised version of that, not a sequel to it. And since several years have passed, some of the passion of the original may have gone, being replaced (I hope) with more careful discussion.

1. INTRODUCTION

Artificial Intelligence is full of 'toy problems': small, artificial axiomatisations or puzzles designed to exercise the talents of various problem-solving programs or representational languages or systems. The subject badly needs some non-toy worlds to experiment with. In other areas of Cognitive Science, also, there is a need to consider the organization of knowledge on a larger scale than is currently done, if only because quantitatively different mental models may well be qualitatively different.

In this document I propose the construction of a formalisation of a sizable portion of common-sense knowledge about the everyday physical world: about objects, shape, space, movement, substances (solids and liquids), time, etc. Such a formalisation could, for example, be a collection of assertions in a first-order logical formalism, or a collection of KRL units, or a microplanner program, or one of a number of other things, or even a mixture of several. It should have the following characteristics.

(1) Breadth. It should cover the whole range of everyday physical phenomena: not just the blocks world, for example.

Since in some important sense the world (even the everyday world) is infinitely rich in possible phenomena, this will never be perfect. Nevertheless, we should <u>try</u> to fill in all the major holes, or at least identify them.

It should be reasonably detailed. For example, such aspects of a block in a block-world as shape, material, weight, rigidity and surface texture should be available as concepts in a blocks-world description, as well as support relationships.

(ii) <u>Density</u>. The ratio of facts to concepts needs to be fairly high. Put another way: the units have to have <u>lots</u> of slots. Low-density formalisations are in some sense trivial: they fail to say enough about the concepts they mention to pin down the meaning of their symbols at all precisely. Sometimes, for special purposes, as for example in foundational studies, this can be an advantage: but not for us.

(iii) <u>Uniformity</u>. There should be a common formal framework (language, system, etc.) for the whole formalisation, so that the inferential connections between the different parts (axioms, frames,...) can be clearly seen, and divisions into subformalisations are not prejudged by deciding to use one formalism for one area and a different one for a different area.

I (still) believe that a formalisation of naive physics with these properties can be constructed within a reasonable time-scale. The reasons for such optimism are explained later. It is important however to clearly distinguish this proposal from some others with which it may be confused, because some of these seem to be far less tractable.

2. WHAT THE PROPOSAL ISN'T

(a) 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. It is tempting to make such demonstrations from time to time. (They impress people; and it is satisfying to have actually made something which work, like building model railways; and one's students can get Ph.D.'s that way.) But they divert attention from the main goal. In fact, I believe they have several more dangerous effects. It is perilously easy to conclude that, because one has a program which works (in some sense), its representation of its knowledge must be more or less correct (in some sense). Now this is true, in some sense. But a representation may be adequate to support a limited kind of inference, and completely unable to be extended to support a slightly more general kind of behavior. It may be wholly limited by scale factors, and therefore tell us nothing about thinking about realistically complicated worlds. Images as internal pictures and the STRIPS representation of actions by add and delete lists are two good examples. I suspect that the use of state variables to represent time is another. Such representational devices are traps, tempting the unwary into dead ends where they struggle to overcome insurmountable difficulties, difficulties generated by the representation itself. I now believe, although I know this view is very controversial, that the famous frame problem is such a difficulty: an apparently deep problem which is largely artifact.

I emphasize this point because there is still a prevailing attitude in AI that research which does not result fairly quickly in a working program of some kind is somehow useless or, at least, highly suspicious. Of course implementability is the ultimate test of the validity of ideas in AI, and I do not mean to argue against this. But we must not be too hasty.

This is no more than a reiteration of John McCarthy's emphasis, since the inception of AI as a subject, on the importance of representational issues (McCarthy 1957, McCarthy & Hayes 1969). In 1969, McCarthy proposed the "Missouri Program", which would make no inferences of its own but be willing to check proposed arguments submitted to it: a proof checker for common sense. Those who find it repugnant to be told to ignore programming considerations may find it more congenial to be urged to imagine the project of building a proof <u>checker</u> for naive physics.

(b) It is <u>not</u> proposed to develop a new formalism or language to write down all this knowledge in. In fact, I propose (as my friends will have already guessed) that first-order logic is a suitable basic vehicle for representation. However, let me at once qualify this.

I have no particular brief for the usual syntax of first-order logic. Personally I find it agreeable: but if someone likes to write at all out in KRL, or semantic networks of one sort or another, or OMEGA, or KRYPTON, or what have you; well, that's fine. The important point is that one <u>knows</u> <u>what</u> <u>it</u> <u>means</u>: that the formalism has a clear <u>interpretation</u> (I avoid the word 's*m*nt*cs' deliberately). At the level of interpreta-

tion, there is little to choose between any of these, and most are strictly weaker than predicate calculus, which also has the advantage of a clear, explicit model theory, and a well-understood proof theory.

I have pointed out elsewhere (Hayes 1977, 1978) that virtually all known representational schemes are equivalent to first-order logic (with one or two notable exceptions, primarily to do with nonmonotonic reasoning). This is still true in 1983, but I should perhaps emphasise that care is needed in making comparisons. First, in claiming equivalence, one is speaking of representational (expressive) power, not computational efficiency. Given a simple "dumb" interpreter (i.e. a "uniform" theorem-prover), these may be at odds with one another. The moral is that simple, dumb interpreters are a bad idea, and interpreters should be sensitive to 'control' information, meta-information about the inferential process itself. This idea brings its own representational problems. I am not arguing that these should be ignored. On the contrary, they raise some of the most important questions in AI. But until we have some idea of the sorts of inferences we might want to control, speculation on the matter is premature. Second, in making comparisons between systems one must exercise care. Many "computational" systems have invisible, buried, assumptions about their domain, not explicitly documented in publications, which must be rendered explicit in a logical axiomatization.[1] Third, the use of logic imposes almost no restrictions on the kinds of thing about which we wish to speak: sequences of actions or views of a room or

plans or goals, etc., are all perfectly fine candidates. One must not let lack of imagination in axiomatising lead one to conclude that logical formalisms are weaker than some of the more superficially baroque systems which AI has devised. (In particular, first-order logic can be taken to quantify over some properties, functions and relations and still be essentially first-order. What makes it higher-order is when its quantifiers have to range over $\underline{all}^2$ properties, functions and relations, a condition which cannot be enforced without something like a rule of λ-abstraction or a comprehension schema.)

Finally, let me emphasize that idiosyncratic notations may sometimes be useful for idiosyncratic subtheories. For example, in sketching an axiomatic theory of fluids (this volume) I found it useful to think of the possible physical states of fluids as being essentially states of a finite-state machine. This summarizes a whole lot of lengthy, and rather clumsy, first-order axioms into one neat diagram. Still, it _means_ the same as the axioms: first-order logic is still, as it were, the reference language. It is essential that there be some standard reference language in this way, so that the different parts of the formalism can be related to one another.

(c) It is not proposed to find a philosophically exciting reduction of all ordinary concepts to some special collection of concepts (such as sets, or Goodmanesque "individuals", or space-time points, or qualia.) Maybe some such reduction will eventually turn out to be possible. I think it extremely unlikely and not especially desirable, but whether or not it is, is not the present issue. _First_ we need to formalise the naive

worldview, using whatever concepts seem best suited to that purpose -- thousands or tens of thousands of them if necessary. Afterwards we can try to impose some a priori ontological scheme upon it. But until we have the basic theory articulated, we don't know what our subject matter is.

Now, this is not to say that we should not exercise some care in avoiding unnecessary proliferation of axioms, or some aesthetic sensibility in designing axioms to give clean proofs and to interact as elegantly as possible. But these are matters of general scientific style, not ends in themselves.

3.   THEORIES, TOKENS AND CLOSURE

Let us imagine that a NP formalisation exists. It consists of a large number of assertions (or: frames, scripts, networks, etc.) involving a large number of relation, function and constant symbols (or: frame headers, slot names, node and arc labels, etc.. From now on I will not bother to reemphasise these obvious parallels). For neutral words, let us call these formal symbols tokens, and the collection of axioms the theory (in the sense of 'formal theory' in logic, not 'scientific theory' in history of science).

The success of a NP theory is measured by the extent to which it provides a vocabulary of tokens which allow a wide range of intuitive concepts to be expressed, and to which it then supports conclusions mirroring those which we find correct or reasonable. People know, for example, that if a stone is re- leased, it falls with increasing speed until it hits something, and there is then an impact, which can cause damage if the velocity is high. The theory should provide tokens allowing one to express the concept of releasing a stone in space. And it should then be possible to infer from the theory that it will fall, etc.: so there must be tokens enabling one to express ideas of velocity, direction, impact, and so on. And then these same tokens must be usable in describing other kinds of circum- stance, and the theory draw the appropriate conclusions there, and so on. We want the overall pattern of consequences produced by the theory to correspond reasonably faithfully to our own

intuition in both breadth and detail. Given the hypothesis that our own intuition is itself realised as a theory of this kind inside our heads, the NP theory we construct will then be equipotent with this inner theory.

More subtle tests than mere matching against intuition might be applied to an NP theory. Consequences which are very obvious should have shorter derivations then those which require some thought, perhaps. If in proving p from q the theory must make use of some concept token, perhaps psychologists can devise an experiment in which the "activation" of that concept can be tested for, while people are deciding whether or not q, given p. Pylyshyn (1979) discusses ways in which intermediate psychological states might be investigated: I will not discuss them further here, but focus instead on questions connected with getting a theory constructed in the first place.

The practical task of building such a theory begins with some 'target' concepts and desired inferences. Take the familiar example of formalising a world of cubical wooden blocks on a flat table, with the goal of being able to reason about processes of piling these into vertical stacks and rearranging such piles by moving blocks from place to place: the familiar "blocks world". Notice that we have put quite a constraint on what inferences we are interested in. An actual tabletop of blocks admits of many more interesting and complicated activities; building walls and pyramids, pushing blocks around horizontally, juggling, etc., but we deliberately exclude such matters from consideration for now.

I will go through this toy world in detail, in order to illustrate some general points. It is not intended as a serious

exercise in Naive Physics. First, we obviously need the concept
of block (a predicate Block(b)), and there will be several states
of the little universe as things are moved, so we also need that
concept (State(s)). A block will be on some other block or on
the table in every state (On(b,c,s), and the name table): four
tokens so far, and now we can write some axioms, such as:

    (i) $\forall$ s,b: State(s) & Block(b): $\supset$

        ($\exists$c. Block(c) & On(b,c,s)) $\lor$ On(b,table,s)

(We could have done it differently: for example, a function
below(b,s) instead of the relation On, so that On(b,c,s)
translates into below(b,s) = c.   Or a function above(b,s), with
the obvious meaning, and a constant, air, so that above(b,s) =
air corresponds to: $\neg\exists$c. On(c,b,s), and being careful never to
apply above to the table. We could have decided not to use
states at all, but to have thought of each block as having a
temporal history. No doubt other variations are possible.   In
the future, I will -- to save paper and to improve readability --
omit such antecedents as Block(b) and State(s) from formulae. It
is straightforward to enrich the logic to a many-sorted logic in
which this omission is syntactically normal. The concepts are
there, though, and need inferential machinery of one kind or
another, so they should be shown in the "reference language".)

    Now, to describe change we need the idea of a state-
transition. There are several ways to do this. We could have a
relation Next(s,t) between states, for example, or a function
next(s), corresponding to the intuitive feeling that one moment

follows another, and there is always a unique next thing that will in fact happen (que sera, sera). Or we might say that, since we are talking about actions, and there are usually several things one <u>might</u> do in a given situation, so there are several different next-states. This leads to McCarthy's idea -- now standard -- of actions as state-to-state functions. We might have actions pickup(b,s) and putdown(b,s), for example. The result of picking up b is a state in which b is no longer on anything but rather is held in the hand:

(ii) $\underline{Held}(\underline{b},pickup(\underline{b}))$

(iii) $\underline{Held}(\underline{b},\underline{s}) \supset \forall x . \neg \underline{On}(\underline{b},\underline{x},\underline{s})$

We must now modify (i) by adding $\underline{Held}(\underline{b},\underline{s})$ as a third possibility. The result of putdowning on <u>b</u> is that whatever is held gets to rest on <u>b</u>; provided of course there is nothing there already. To make this neater, let's define <u>Clear</u>:

(iv) $\underline{Clear}(\underline{x},\underline{s}) \supset . \forall c . \neg \underline{On}(\underline{c},\underline{x},\underline{s}) \lor \underline{x}=\underline{table}$

Then we can say:

(v) $\underline{Held}(\underline{b},\underline{s})$ & $\underline{Clear}(\underline{c},\underline{s}) \supset \underline{On}(\underline{b},\underline{c},putdown(\underline{c},\underline{s}))$

(This still doesn't explain what putdown(<u>c</u>,<u>s</u>) is like if nothing is <u>Held</u> in <u>s</u>. We might decide there are two sorts of states, those in which the hand is holding something and those in which it is empty, and insist that putdown applies only to the former. Or we might just say that:

(vi) $(\forall \underline{x} . \neg \underline{Held}(\underline{x},\underline{s})) \supset putdown(\underline{c},\underline{s}) = \underline{s}$)

We can now begin to see how the desired kinds of conclusion might follow. If we know that A is on C on the table and B is on the table and A and C are clear, then we can infer from (ii) that after a suitable pickup, A is held. Unfortunately, we can't conclude that B is still clear: C may have jumped onto it, as far as our axioms are concerned. (Consider a world of jumping blocks, or stackable frogs, in which every time one is lifted, the one beneath hops onto a different block. This is a possible world, and all five of the axioms are true in it. So, nothing that they say rules this possibility out.) This is a tiny illustration of the notorious frame problem (McCarthy & Hayes 1969). We need to say that during a pickup of a block b, no other On relations change.

Now, for the first time, we don't need to introduce any new tokens. We have a rich enough vocabulary at hand to state our axiom:

(vii) $\underline{On}(\underline{b},\underline{c},\underline{s}) \supset \forall \underline{d}.\ \underline{On}(\underline{b},\underline{c},\underline{pickup}(\underline{d},\underline{s})) \veebar \underline{b}=\underline{d}$

Here, is exclusive-or, so that if b is not d, then $\underline{On}(\underline{b},\underline{c})$ must still be true in $\underline{pickup}(\underline{d},\underline{s})$; and we are sure that $\neg\,\underline{On}(\underline{b},\underline{c},\underline{pickup}(\underline{b},\underline{s}))$ under any circumstances. Notice that the block picked up might itself carry others, and they go right along with it.

Given (vii), we can quickly conclude that B is still clear and still on the table), so we can now putdown onto it and have a state in which A is on B -- no longer clear, by (iv) -- and C is clear...well, not quite, since putting down might yet disturb things. But we can fix this with an even simpler frame axiom:

(viii) $\underline{On}(\underline{b},\underline{c},\underline{s}) \supset \underline{On}(\underline{b},\underline{c},\underline{putdown}(\underline{d},\underline{s}))$

and we can now discuss states reached by picking up and putting down things all over the place, as we desired. Given a sufficiently complete description of a layout of blocks, and a goal of some other configuration, then if there _is_ a sequence of block movements which get us from the former to the latter, then this theory will show that there is.

For some time now we have not needed to introduce any other tokens. We can do the changes by adding or modifying axioms, working entirely in the given vocabulary. This collection of tokens: (_block_, _table_, _state_, _on_, _held_, _pickup_, _putdown_), is enough to work with. Alternative worlds can be constructed within it. It is a large enough collection to support axioms describing general properties of the universe we have in mind, and descriptions of particular worlds in enough detail to allow the sorts of conclusion we wanted to be inferred. No subset will do the job, as we have seen:[3] but this is just enough to let us say what needs saying. We have reached what might be called a conceptual closure. This phenomenon is familiar to anyone who has tried to axiomatise or formalise some area. Having chosen one's concepts to start on, one quickly needs to introduce tokens for others one had not contemplated, and the axioms which pin down their meanings introduce others, and so on: until one finds suddenly there are enough tokens around that it is easy to say enough "about" them all: enough, that is, to enable the inferences one had had in mind all along to be made.

This sort of closure is by no means trivial. Suppose we had tried to use next(s), following the idea that world-states are, after all, linearly ordered; then it becomes quite hard to achieve. We can say that a block may stay where it is, or become picked up:

(ix) $\underline{On}(\underline{b},\underline{c},\underline{s})$ & $\forall \underline{x}. \neg \underline{Held}(\underline{x},\underline{s}). \supset$

                         $\underline{On}(\underline{b},\underline{c},\underline{next}(\underline{s})) \lor \underline{Held}(\underline{b},\underline{next}(\underline{s}))$

and we can insist that only one is held at once:

(x) $\underline{Held}(\underline{b},\underline{s})$ & $\underline{Held}(\underline{c},\underline{s}) \supset \underline{b}=\underline{c}$

But putting down is more difficult. If we say

(xi) $\underline{Held}(\underline{b},\underline{s})$ & $\underline{Clear}(\underline{c},\underline{s}) \supset \underline{On}(\underline{b},\underline{c},\underline{next}(\underline{s}))$

then the held block has been put down into every clear space. We want to say that the held block is put down in one of the potential putdown sites:

(xii) $\underline{Held}(\underline{b},\underline{s}) \supset . \exists \underline{c}. \underline{Clear}(\underline{c},\underline{s})$ & $\underline{On}(\underline{b},\underline{c},\underline{next}(\underline{s}))$

But we now have no way of inferring that the held block can actually be placed in any particular clear place. This axiom is consistent with a world in which blocks can be placed only on the table, for example, or in which blocks are always released from on high and falleth gently upon some random stack or other. There is no way within this vocabulary to describe one possible future state's properties as distinct from those of a different possible future state. We have no way of stating the properties we need: closure eludes us. It can be achieved, but only by

bringing possible futures in by the back door.

Our theory, though closed, is by no means perfect. As stated, it can support all the inferences we had in mind. Unfortunately, it can also support some others which we didn't have in mind. For example, nothing in the axioms so far prevents two successive pickups, giving a handful of blocks (or, somewhat less plausibly, a handful of towers of blocks). This would be fine, except that (v) has it that anything held is deposited by a putdown, thus leaving several blocks on one: but they were supposed to be all the same size. The neatest way to fix this is to modify (ii), say as follows:

(xiii) $(\forall x. \neg \underline{Held}(\underline{x},\underline{s})) \supset \underline{Held}(\underline{b},\underline{pickup}(\underline{b},\underline{s}))$

We can also insist that only single blocks are picked up by adding $\underline{Clear}(\underline{b},\underline{s})$ as another antecedent condition. Again: if a block is $\underline{Clear}$, then we can pick it up -- its still $\underline{Clear}$ -- and put it down $\underline{on}$ $\underline{itself}$: there's nothing in (v) to prevent this. (Consider a zero-gravity world in which blocks can be released in space, and they then just hang there: and say that in this case the block is $\underline{On}$ itself. Clearly all the axioms are satisfied in this world too.) So to rule this out we need another axiom, and to modify (v) slightly. Finding other such bugs is left as an exercise for the reader.

It is important to bear such negative properties of a formalisation in mind even though they make the formaliser's life more complex. It is easy to overlook them.

4. MEANINGS, THEORIES AND MODEL THEORY

In developing this toy theory I have several times used an example world to show that something we wanted to follow didn't, or that something we didn't want to be true might be. This ability to interpret our axioms in a possible world, see what they say and whether it is true or not, is so useful that I cannot imagine proceeding without it. But it is only possible if there is an idea of a model of the formal language in which the theory is written: a systematic notion of what a possible world is and how the tokens of the theory can be mapped into entities (or structures or values or whatever) in such worlds. We have to be able to imagine what our tokens might mean.

Now this semantic metatheory may be relatively informal, but the more exactly it is defined, the more useful it will be as a tool for the theory-builder. The main attraction of formal logics as representational languages is that they have very precise model theories, and the main attraction of first-order logic is that its model theory is so simple, so widely applicable, and yet so powerful.

A first-order model is a set of entities and suitable mappings from tokens to functions and relations, of appropriate arity, over it.[4] Any collection of things will do: for example, for our blocks world, I could take the collection of papers on my desk, and interpret On to be the relation which holds between two pieces of paper when one partially or wholly overlaps the other, and pickup to be the action of picking up, and so on. (In fact, this isn't a model, because my desk is too crowded: axiom (v) is

false. But it would be if I tidied my desk up.)

This is very satisfying, since we have found a model which is very close to the original intuition. But there are other models. Consider a table and a single block and the two states, one -- call it A -- with the block on the table, and the other -- call it B -- with the block held above the table. Let _pickup_ and _putdown_ denote the functions (A → B,B → B) and (B → A,A → A) respectively, let _Held_ be true just of the block in state B, and let _On_ be true just of the block and the table in state A. All the axioms are true, so this is a possible world. This one is much simpler than my desk, and its existence shows that the axioms really say rather less than one might have thought they did: specifically, they say nothing about _how_ _many_ blocks or states there are, or about the direction of time's arrow.

One can find other very simple models, for example models made of dots being moved on a screen -- so the theory says nothing about the 3-dimensionality of the world.

This illustrates how the existence of a model theory for our formal language is not just a methodological convenience. It tells us what our formalisations mean and hence, what they don't mean. We may think that we have captured some concept in a theory, but unless the theory is sufficiently rich to guarantee that _all_ its models reveal the kind of structure we had in mind, then we are deluded: _a token of a theory means no more than it means in the simplest model of the theory._

Returning to methodology for a moment, a crucial property of this way of characterising meaning is that it transcends syntactic and operational variations. A given theory might be

realised operationally in innumerable ways. Even ignoring heuristic 'control' issues, we have such variations as natural deduction rules, or semantic tableaux or Hilbert-style axiomatisations. We can make the theory look like a semantic network or a collection of frames or MOPS or any one of innumerable other variations. None of these variations will give the theory an ounce more expressive power. None of them could ever make good a representational inadequacy of the theory. It is easy to lose sight of this basic and uncomfortable truth. Thinking model-theoretically helps us to keep it in mind.

It also gives us a powerful theoretical tool. I mentioned earlier that defined concept tokens, such as Clear, added no real expressiveness to a theory. This seems kind of intuitive once it is pointed out, but it has a quite conclusive model-theoretic statement (Beth's definability theorem) which completely settles the matter, and frees up time for more productive discussions.

An objection to the idea of models goes as follows. Any particular formalisation or implementation consists entirely of the expressions and the inference rules or procedures which manipulate them. The idea of a model, and the mappings which relate expressions to denotations, etc., are just metatheorists ideas, imposed from without. But we could have a different model theory for the same formal language, and declare that this seman-tic theory assigned meanings to the formal symbols. (E.g., see D. Israel, this volume.) And who is to say which of the many possible semantic theories is the right one?

But the relationship between a model theory and the (purely formal) inference rules or procedures attached to the formal language is not arbitrary in this way. Each model theory sanctions certain inferences (the ones that preserve truth in those models) and not others. And, sometimes, we also get the converse, viz., if some assertion is time in all those models, then the rules will indeed eventually declare it so. This is the content of the completeness theorem for a formal language. We should treasure completeness theorems: they are rare and beautiful things. Without them, we have no good justification for our claims that we know how our theories say what we claim they say about the worlds we want them to describe. To emphasise this, consider enriching the formal language by introducing a new kind of symbol, say a quantifier M which I claim means 'most' so that MxP(x) means P is true of most things. I can easily give a model theory: MxP(x) is true in a model just when P is true of more than half the universe (with a little more subtlety for infinite domains, but let that pass). I can claim this, but the claim is premature until I can describe some mechanism of inference which captures that interpretation, generating all the inferences which it justifies and none which it refutes. And this might be difficult. For some model theories we know it is impossible.

A model theory can determine the actual meaning of the logical symbols of the formal language, but it does not determine the actual meaning of the tokens. The only way to do that is by restricting the set of possible models of the theroy, for example by adding axioms. All we can say of a token is that in this model

it means this, in that one it means that. There is no single 'meaning' of a formal token (unless there is only a single model): we cannot point to something and say, that is the meaning.

We might restate the goal of building a formal theory as being that of ensuring that all the models of the theory are recognisable as the kind of possible world we were trying to describe, so that in each one, each token denotes what it should. But this notion of meaning raises a wellknown philosophical spectre, a second objection to a model-theoretic view of meaning. For no model theory can specify what kinds of entity constitute the universes of its models. It refers only to the presence of functions and relations defined over a set, not to what the set is a set of. And we could always make our universes out of entirely unsuitable things, in particular the tokens themselves.

Suppose we have a 'suitable' model of a theory. Make a ghost model as follows. Let each name denote itself. Every token which should denote an operation on things, interpret it rather as an operation on the names of things, whose result is the expression which would have referred to the thing got by performing the operation on the things named, so that for example a unary function symbol f denotes the function on expressions which takes the expression 'e' to the expression '$f(e)$', '$g(h(a))$' to '$f(g(h(a)))$', and so on. And interpret each relation symbol as that relation on expressions which is true when the relation is true of the thing named by the expressions

in the 'suitable' model: so that P denotes the predicate which is true of the symbol 'a' just if 'P(a)' is true in the first model. In general, whenever you need to decide a question of fact, go and check in the 'suitable' model to see what its facts are, and use those.

There is one of these ghostly (Herbrand) models for every model, and it makes exactly the same axioms true. So there could be no way of adding axioms (or frames or scripts or demons or MOPS or anything else, just to re-emphasize the point) which could ensure that all a theory is talking about might not be its own symbols.

This is an important point, considered as a criticism of a theory of meaning. Indeed, no formal operations, no matter how complex, can ever ensure that tokens denote any particular kinds of entity. There are, I think, three ways in which tokens can be attached to their denotations more rigidly (so to speak). One; if the token is itself in a metatheory of some internal part of the theory, then the connection can simply be directly made by internal, formal, manipulations. Formally, these are "reflection principles", or rules of translation between a language and its metalanguage.

Two; if the theory is in a creature with a body -- a robot, like us -- then some of the tokens can be attached to sensory and motor systems so that the truth of some propositions containing them is kept in correspondence to the way the real world actually is. These tokens -- they might include the concept of vertical connected to the inner ear, and those of a whole intricate theory of lighting and surfaces and geometry and texture and movement

connected with visual perception, and a whole other collection associated with proprioceptive awareness of the body's position in space -- have a special status. We might say that the body's sensorimotor apparatus _was_ the model theory of this part of the internal formalisation.

Three, tokens could be attached to the world through language. Again, let the theory be built into a physical computer, one without senses, but with a natural-language comprehension and production system. The tokens of the internal theory are now related to English words in the way we expect, so that the deep semantic meaning of a sentence is a collection of axiomatic statements in the formalism. Such a system could talk about things to other language-users and could come to learn facts about an external world by communicating with them. Assuming that _their_ beliefs and conversations really were about things -- that they managed to actually refer to external entities -- then I think we would have no reason to refuse the same honor to the conversing system.

These matters require and deserve fuller discussion elsewhere. But I suggest that for the purposes of developing a naive physics, this whole issue can be safely ignored. We can take out a promissory loan on _real_ meanings. One way or another, parts of our growing formalisation will have eventually to be attached to external worlds through senses or language or maybe some other way, and ghost models will be excluded. We must go ahead trying to formalise our intuitive world; paying attention indeed to the complexity and structural suitability of our

models, but not worrying about what sort of stuff they are made from.

We have then to be ready to repay the loan, by looking out for areas of axiomatisation where the tokens might be attachable to perceptual or motor or linguistic systems. For example, ideas connected with time must make some contact with our internal 'clocks' of various sorts. Much of our inituitive knowledge of force and movement comes from _what_ _it_ _feels_ _like_ when we push, pull, lift and move. Much of our knowledge of 3-space is connected with how things _look_; and so on.

5. DISCOVERING INTUITIONS AND BUILDING THEORIES

We have been assuming all along that we are able to interpret tokens of the theory in intuitive terms. But this assumes that we can identify our own intuitive concepts sufficiently clearly to assign them to tokens. In practice, building axiomatic theories is in large part an exploration and clarification of our own intuitions. Just as professional grammarians tend to acquire an astonishingly acute sense of exactly which syntactic constructions are acceptable to a native speaker, so naive physicists will need to develop an acute sense of intuitive reasonableness of descriptions of the everyday physical world. It is not at all an easy thing to do.

Consider the earlier toy blocks-world example. It might be argued that here is a small theory with complete conceptual closure. But it is closed only with respect to the very limited range of inferences we required initially; this is exactly what makes it a toy theory. Try to expand it to deal with our own ideas of putting things on things. We have the token On: what exactly did that mean? It had a component of pure geometry, referring to the spatial arrangement of the blocks. It also seemed to have some idea of support contained within it: if A is on B, then B is holding A up; B is the reason why A isn't falling, it is bearing A's weight. Now these are very different ideas. For example, the geometric On is asymmetrical (nothing is on anything which is on it -- although it doesn't seem that this should be an axiom so much as a consequence of some more basic

spatial theory), but the support On can be, e.g., two long blocks leaning on one another. They come together here in that the geometric On implies the support On, because blocks are rigid and strong, so they will bear weight without deforming or breaking. And this is because the stuff they are made of has these properties. To emphasize the separateness of these two ideas, imagine the alternative possible world with no gravity. The geometry is unchanged, but the 'support' idea is absent. So they must have distinct subtheories.

Both concepts are linked to clusters of others which we have not yet begun to formalize. The experience of doing so may well sharpen our sense of what the concept is, perhaps separating it out further into several slightly (or very) different ideas, each requiring its own axiomatic connections to the rest of the theory.

We have taken a proposed concept and seen it as a blend of two distinct components. As well as this analytic "division" of concepts there is what we might call a process of "broadening"; extending the range of a concept, trying it out in other areas where it seems natural. For example, imagine four blocks arranged in a compact square on the table, with adjacent faces in contact (the very fact that you can do that says a lot about the richness of the spatial-geometry part of our internal theories) and place a fifth block neatly on top, in the center. What is this block on? We might say on each of the other blocks, but this is a very different notion (e.g. pick up one of the lower blocks). Perhaps it is on the set of the four blocks ... but a set hardly seems the kind of thing that can bear weight, and

anyway only some sets will work. Perhaps we should abandon the notion of _on_ altogether in this case in favor of some other, more subtle, relationship between the blocks. But it seems intuitively clear that the top block _is_ on something, in much the same way that it could be on one block.   The only reasonable conclusion, I believe, is that the fifth block is indeed on a (single) thing, which is made up of the four other blocks. By arranging them thus in a compact square, one has created a new object; we might call it a platform. (If someone points to it and asks; what is _that_?, the question is quite intelligible: there is some _thing_ there.   One might of course answer: nothing, its just four blocks.) So blocks can be on other things than blocks and tables. Its the same concept, but using it in a different situation forces a reevalation of what can be said about it.

We need to be able to state some criterion of put-on-ability, which seems to be having a firm horizontal surface. But now we have a new concept, that of a surface. This requires more axioms to relate it to existing concepts, and these in turn introduce other concepts (edge, side of a surface, direction, adjacency, contact, the object-surface relation, etc.: see 'liquids', this volume, for a first attempt at such a list) and these require more axioms, each typically introducing other concepts, and so on. Conceptual closure becomes much harder to achieve: perhaps impossible to achieve completely.

This is what typically happens when one extends the scope of a concept. Closure is fragile, sensitive to the demands placed on the theory.   Toy theories achieve it only by having very

restricted demands placed on them. In developing naive physics we expect far more of the theory, forcing it to be larger and making closure more remote. There is a constant tension between wanting a closed theory and wanting to pin down the meanings of tokens as precisely as we can: between closure and breadth.

This example illustrates an important and basic fact about the enterprise of knowledge representation. We want breadth and density: but you can't have the density without the breadth. If we want the theory to say a _lot_ about a concept, the only way to do so is to relate that concept to many others. If there are many axioms in the theory which contain a certain token, there must _ipso facto_ be many other tokens to which it is axiomatically related. It is exactly this, being tightly caught in a dense web of inferential connections to other parts of the theory, which gives a token meaning, by cutting out unwanted implausible models. And this is what we want, since the goal of the axiomatising enterprise is to produce a theory from which we can rapidly draw the many conclusions corresponding to our intuitions, and this inferential richness goes along with model-theoretic constraint.

It is easy to find other things wrong with the toy blocks world: it was always just a toy, in any case, and we will now abandon it. But its limitations illustrate a serious general problem of how to get naive physics done.

A completed theory would be huge (a guess: between 10^4 and 10^5 tokens). It would be conceptually closed,[5] but it seems overwhelmingly likely that no reasonably sized subtheory of it will be. Such a subtheory would be completely isolated from the

rest of the theory: the meanings of its tokens would not be affected by the way in which the other axioms imposed interpretations on the rest of the tokens.

It seems much more likely (it is in any case the most conservative assumption) that the whole theory is bound together, so that the meaning of any token depends on all of the rest of the theory. But then how can we judge the correctness or suitability of part of such a theory? Since at any intermediate stage of theory construction there will be tokens not yet axiomatised, the process of formalising those concepts may force changes in their correspondence to intuition and these changes might require our earlier partial theories to be rewritten. The toy blocks world's concepts came apart and its axioms became inappropriate to the new meanings, when we divided it into separate geometric and physical components, for example. Anyone who has tried to expand the scope of an existing representation will recognize the problem, but the methodology being urged here seems to preclude all the usual solutions.

One response is to proceed by enlarging the toy problems. On this approach we will work on progressively more ambitious subtheories, but always with a clear boundary on the kinds of inferences which the theory is expected to support. This approach is however very dangerous, since it can get caught in conceptual traps, as noted earlier. A technique might work well in a limited domain, and be applicable - with increasing difficulty - to a wider and wider range of phenonena, but be ultimately wrong. It is perilously easy to go on putting off

consideration of the examples which clearly demonstrate its futility: one always plans to get to those later. Our toy blocks world embodies several such errors, notably the use of state - state functions to denote actions (completely unusable when several things are happening at once: see Section 6 below).

Another response is to search for a small kernel theory of basic concepts, to which all others can be reduced by suitable definitions. Put another way: suppose we had a finished naive physics and eliminated all tokens which were explicitly defined in terms of others (as clear was in the blocks world), kept on doing this to the limit, and looked at what was left. This reduced theory must be conceptually closed, since the original one was: call it the kernel. Now, perhaps this is a smallish theory (less than a thousand tokens) so that to get it all done would be a feasible project. Filling in the rest can then be done piecemeal as needed, since adding definitions of new tokens does not affect the meanings of the old tokens; there can be no forced revision of the kernel axiomatization. This is the "semantic primitives" idea exemplified in the work of R. Schank and Y. Wilks.

It is worth pointing out that such a small kernel theory supporting a much larger theory by means of mere definitions does exist. It is axiomatic set theory, and it supports virtually the whole of pure mathematics. We have had 60 years to get used to the idea, but it is incredible that such an audacious program should have so nearly succeeded: a tiny theory (2 tokens and perhaps 8 axioms: details vary) enables one to define a large number of mathematical concepts, and then provides enough

inferential power that the properties of these things follows from their definitions. The induction principle for the integers, for example, is not an axiom, but a theorem whose truth can be established within set theory.

Maybe such a small kernel can be found for our conceptual theory, but I very much doubt it. It seems a priori implausible that our knowledge of the rich variety of the everyday world could be merely a collection of lemmas to some small set of concepts. And there is a more technical objection, borne out by experience with schemes of scientific primitives. To pin down a concept exactly requires a rich theory and hence a large theory: exactness entails density which entails breadth, as we noted earlier. It follows that a small theory which is conceptually closed and yet has a wide scope cannot be detailed. The concepts it discusses must be at a high level of generality not very tightly constrained by the theory. But then, if all else we have are definitions, we will never be able to get at the details. As Wilks (197) says, no representation in terms of primitives can be expected to be able to distinguish between hammers, mallets and axes. But we must, somehow.

A third response, less idealistic but I think inevitable, is to accept the problem as real and find ways to live with it. We must build theories which are only partially closed. Some tokens will not yet have their meanings axiomatically specified: they will represent directions for future investigation. We will, indeed, always be in danger of having later theory construction come back and force an alteration in our present work, perhaps

scrapping it entirely. The best we can hope for is to develop a good sense of style and scope in choosing groups of concepts and in formulating their subtheories.

Breadth seems to be crucial. If a concept makes intuitive sense in a wide variety of circumstances, but its candidate theory somehow presupposes a more limited framework, then something is wrong. Either the concept has several parts or cases, one of which is provisionally captured by the theory: or, more likely, the theory is limited by some inappropriate restriction (e.g. that blocks can be put only on other blocks) and needs to be recast in different terms. Applying this breadth criterion as a heuristic guide when building theories is what most clearly distinguishes this from the toy-worlds approach. Sometimes one has to accept a limitation for no better reason than that one can see no way to make progress without it, but this is to be resisted, rather than taken as a guiding principle.[6]

So far, I have assumed that concepts have been initially identified by no more than careful introspection. Other more objective and disciplined ways are also available. Detailed examination of the meanings of English spatial prepositions (Herskowitz 198...) provides many clues. Driving introspection deeper by sensitive interviews (Gentner 1983) can uncover the outlines of whole inner theories. Showing subjects simplified physical situations (or tricking them with excruciatingly realistic ones: Howard) and finding their intuitive predictions can clearly reveal centrally important concepts (such as "impetus" (McCluckey 1983)).

Many parts of the psychological and linguistic literature are ripe with clues. But one has to exercise great care. It is very difficult to make a <u>direct</u> connection between any aspect of overt behavior and any small part of the conceptual theory, if the present account is anything like correct. Single concepts will not usually emerge as English words, for example. Natural language is for communication, the internal language of thought is for thinking - in our model, inference-making in a highly parallel computer. These are vastly different requirements and so the languages can be expected to be very different. A communication language must be compact (since it has to be encoded as a time sequence, and time is short) but it can afford to be highly context sensitive in the way it encodes meaning (since the recipient is a powerful processor and shares a great deal of the context): neither applies to the internal language.

A word like "in" seems to expand into a whole complex of ideas when examined in detail: we must attempt to build a coherent formal theory of these before making judgements about the appropriateness or otherwise of the expansion, since tokens in isolation are meaningless (and they can <u>seem</u> to be meaningful: see McDermott 198),

6. CLUSTERS

Concepts will not be evenly spread throughout a theory. Some groups of concept-tokens will have many tight axiomatic connections within the group, relatively few outside. Think of a graph with tokens as nodes, linked by an arc if there is an axiom containing both of them: call it the axiom-concept (a-c) graph. Then this graph, while connected, will have some areas more densely connected than others. Call such a collection a cluster. Our job as theory-builders is made easier if we can identify clusters: these are as close as one can get to isolated subtheories.

Identifying clusters is both one of the most important and one of the most difficult methodological tasks in developing a naive physics. I think that several serious mistakes have been made in the past here. For example, causality is, I now tend to think, not a cluster: there is no useful, more-or-less self-contained theory of causality. 'Causality' is a word for what happens when other things happen, and what happens, depends on circumstances. If there is liquid around, for example, things will often happen very differently from when everything is nice and dry. What happens with liquids, however, is part of the liquids cluster, not part of some theory of 'what-happens-when'. This is not to say that the concept of causality is useless, but that it is an umbrella term for a large variety of particular relationships, each of which has its own detailed cluster of supporting theory, and its meaning is parasitic on theirs. If all you know is that A caused B, about all you can conclude is

that A was before B.

Mistakes like this are hard to overcome, since a large conceptual structure can be entered anywhere. The symptom of having got it wrong is that it seems hard to say anything very useful about the concepts one has proposed (because one has entered the graph at a locally sparse place, rather than somewhere in a cluster). But this can also be because of having chosen one's concepts badly, lack of imagination, or any of several other reasons. It is easier, fortunately, to recognise when one is in a cluster: assertions suggest themselves faster than one can write them down.

A good strategy seems to be to work on clusters more or less independently at first: the meaning of the tokens in a cluster is more tightly constrained by the structure of a cluster than by the links to other clusters. It seems reasonable therefore to introduce concepts, which occur definitively in some other cluster, fairly freely, assuming that their meaning is, or will be, reasonably tightly specified by that other cluster. For example, in considering liquids, I needed to be able to talk about volumetric shape: assuming - and, I now claim, reasonably - that a shape cluster would specify these for me. Of course, their occurrence in the liquids cluster does alter their meaning: our concept of a horizontal surface would hardly be complete if we had never seen a large, still body of water - but the assumption of a _fairly_ autonomous theory of shape still seems reasonable, at least as a working hypothesis.

The rest of this section discusses some likely clusters and

some of the difficulties and issues which arise in formalizing them.

(a) Places and Positions

Consider the following collection of words: inside, outside, door, portal, window, gate, way in, way out, wall, boundary, container, obstacle, barrier, way past, way through, at, in.

I think these words hint at a cluster of related concepts which are of fundamental importance to naive physics. This cluster concerns the dividing up of 3-space in pieces which have physical boundaries, and the ways in which these pieces of space can be connected to one another; and how objects, people, events, and liquids can get from one such place to another.

There are several reasons why I think this cluster is important. One is merely that it seems so, introspectively. Another is that these ideas, especially the idea of a way through and the things that can go wrong with it, seem widespread themes in folklore and legend, and support many common analogies. Another is that these ideas have cropped up fairly frequently in looking at other clusters, especially liquids and histories (see below). Another is that they are at the root of some important mathematics, viz. homotopy theory and homology theory. But the main reason is that containment limits causality. One of the main reasons for being in a room is to isolate oneself from causal influences which are operating outside, or to prevent those inside the room from leaking out (respectively: to get out of the rain, to discuss a conspiracy). A good grasp of what kind of barriers are effective against what kinds of influence seem to

be a centrally useful talent needed to be able to solve the
'frame problem'.

There is another, closely related idea which could be called
aposition (although the meanings of the English words 'place'
and 'position' do not exactly coincide with the two concepts I am
trying to distinguish). A position is a point within a space
defined by some coordinate frame for that space. This need not
necessarily be a Cartesian frame (in fact, it is rarely so), just
some way of referring to parts of the space (such as the back,
center and front of a stage, or a hotel room numbering system).
A position is a place you can be at; a place is a place you can
be in. Places always have boundaries, positions usually do not.
(Although the boundary may not be marked by a physical barrier,
it is there, and there is a clear notion of crossing it and
getting into the place. Territorial animals have the same idea.)
A position in a space is essentially pointlike in that space's
coordinate system (i.e. it has no internal structure), but it may
itself be a place, in which case its interior is a new space with
its own coordinate system defining positions within it. The
internal coordinate system need have no relation to the external
one, even when there is no physical boundary. For example, one
can be in a corner of a room, a place whose orientation is
radially outwards, but the room's natural coordinate system might
be in terms of a back and a front, left and right.

A room in an apartment in an apartment building is a place
which is a position in the interior of a place which is a
position in the interior of a place. To be in the kitchen is to

be _at_ a position in the apartment (so answers the question: where are you?), and to be _in_ the city is to be _at_ a position in the state or country (so also answers the question, _if_ the space being discussed is this larger one).   This mutual nesting of places and positions can get very deep.[7]   Notice that if one place is inside another then it must be a position within the latter.   (After all, it must be somewhere, right?)

To get in or out of a place is to follow a path which must intersect the boundary.   (This is the basic properties of boundaries).   A path must consist of empty space, so if anything can get in or out, then there must be a part of the boundary of a place which is not solid: the door or portal, the _way_ in or out. It follows that a way to prevent entry or exit is to ensure that there are no holes in the boundary of a place.

(b) Spaces and Objects

Places and positions are concerned with space in the large, space to be in.   But there is also a collection of concepts to do with local small-scale space, the space between and around solid objects.   The two interact, if only in that suitable solid arrangements can define places, by being a boundary.   But there seem to be some concepts and difficulties special to the small scale.

For naive physics, vertical gravity is a constant fact of life, so vertical dimensions should be treated differently from horizontal dimensions: 'tall' and 'long' are different concepts. An object's shape is also often described differently (width and length; or depth - from the wall - and width or length along the wall: width if one thinks of the object as being put against the

wall, length if one thinks of it as running along the wall). I suspect - the details have not been worked out - that these differing collections of concepts arise from the reconciliation of various coordinate systems. A wall, for example, defines a natural coordinate system with a semi-axis along its normal.

An important aspect is the relationship of surfaces to solids and edges to surfaces. The different names available for special cases indicates the richness of this cluster: top, bottom, side, rim, edge, lip, front, back, outline, end. Roget's Thesaurus (class two, section two) supplies hundreds more. Again, these are not invariant under change of orientation, especially with respect to the gravity vertical. Such boundary concepts are also crucial in describing the shape of space, and are the basis of homology theory and differential geometry. There is an obvious connection to the notion of place, in that places have boundaries. Let $\triangle$ be the function which defines the boundary of any piece of space: then $\triangle^2 p$ is the boundary of the boundary of p. If there is a gap in the boundary, then $\triangle^2 p$ is the outline of that gap (the door frame, for example). Homology theory takes it as axiomatic that $\triangle^2 = \emptyset$, and studies the algebraic properties of triangulations which divide space into discrete pieces.

Oneconcept which I currently find especially vexing is that of touching. Intuitively, it seems quite clear. Two bodies can touch, and when they do, there is <u>no</u> <u>space</u> between them: this could even be a definition of touching. It is also clear that they do not (usually) merge together or become attached or

unified into on object: each retains the integrity of its bounding surface. And it also seems intuitively clear that the surface of a solid object is part of the object: the surface of a ballbearing is a <u>steel</u> <u>surface</u>, for example. And, finally, the local space we inhabit does seem to be a pseudo metric space (in the technical sense), i.e. there is a (fairly) clear notion of distance between two points. Unfortunately, taken together, these intuitions are incompatible with the basic assumptions of topology , and it is hard to imagine a more general theory of spatial relationships. Briefly, the argument goes: a pseudometric space is normal, which is to say that if two closed sets of points are disjoint, then there are disjoint open sets each containing one of them. (Intuitively, two closed sets cannot touch without having some points shared between them.) But if objects contain their surfaces, then they are closed sets: so they can never touch.[8]

My treatment of surfaces and contact in "liquids" escapes this problem by saying that when objects touch there is an infinitesimally thin layer of space (the "directed surface") between them. This works up to a point, but seems unintuitive and in any case does not address the basic issue, which is that our intuitive local space is, indeed, probably not a topological space.

It is certainly not three-dimensional Cartesian space, which contains such wildly implausible objects as space-filling curves and the Alexander Horned Sphere. Many mathematical intuitions at the basis of geometry and real analysis (from which topology is an abstraction) seem to be at odds with the way we think about

everyday space. The idea of a point is itself one which people with no mathematical training seem to find difficult, or even incoherent.[9] As with many of the pathological constructions, the difficulty seems to arise from taking reasonable intuitions to unreasonable lengths by introducing infinite limits of one kind or another (infinitely small spots, or infinitely thin lines; surfaces which have no thickness at all, yet are actually there, etc. see (h) below). Intuitive space has a definite "grain" to it: when distances get too small, they cease to exist. It is a tolerance space (Zeeman 1962, Poston 1971, Roberts 1970) rather than a topological space.

All of this intricacy came from taking the idea of "touching" seriously, and illustrates again the way in which trying to capture one concept with some breadth of application can force major changes to large parts of the growing theory.

(c) Qualities, quantities and measurements

Many everyday things have some properties which are more intrinsic than others, and might be called the possession of certain qualities. Objects have sizes, weights, colors; spaces have volumes; some objects have heights, others lengths. All of these qualities seem to exist independently of the entities which possess them. We can discuss heights, colors or smells as things in their own right: they form quality spaces. The set of possible heights is a quality space, as is the set of possible flavors.

There does seem to be a general theory of quality spaces. It always makes sense to consider the extent to which two

qualities are alike: the degree of similarity between them. (Even when the answer is trivial, the question is never incoherent.) Thus there seems to always be a notion of "distance" defined on a quality space. Similarly, all quality spaces seem to have a tolerance. If two qualities are very similar, they become indistinguishable. (This may be the basic structure, as every tolerance defines a natural notion of distance between qualities to the smallest number of steps by which one quality can be transformed into the other, each step being 'invisible' under the tolerance. Poston (1972) develops this idea very thoroughly.) Many quality spaces are dense, in the sense that given any two distinct qualities there is a third somewhere between them. (Colors are dense, but smells and flavors aren't, I think.) Some spaces (colors, notably) seem to be structured in terms of a subset of prototype qualities, the others being defined by their distances from the prototypes. Some seem to be naturally n-dimensional, for some small n: other not.

Some quality spaces can be measured; i.e. there are functions (usually more than one) from them to a measuring scale, a linearly ordered set of some kind (e.g. the positive integers, the rational unit interval, the set {small, smallish, medium, tallish, tall}). Such measure functions (feet, meters) induce an order structure on the quality space (but it may not be a strict linear order). We can use this apparatus to talk about quantities: heights and distances are quantities, colors and smells aren't. We can write for example:

meters (height(Bill)) =

feet (height(Bill)) = 5.9

roughly (height(Bill)) = tallish

Notice that we can discuss heights directly, for example by writing

height(Bill) > height(Fred)

where the ordering relation > is that which is induced by the measuring functions feet and meters. (If we used the similar relation induced by the measuring function roughly, then the above would say something like: Bill is clearly taller than Fred.) One remark which may be apposite here is this. It is often argued that 'common sense' requires a different, fuzzy logic. The examples which are cited to support this view invariably involve fuzzy measuring scales or measure spaces. This, I believe, is where fuzziness may have a place: but that is no argument for fuzzy truth-values.

(d) Change, time and histories

The now classical approach to describing time and change, invented first by J. McCarthy (1957), uses the idea of a state or situation (or: world-state, time-instant, temporally possible world,...). This is a snapshot of the whole universe at a given moment. Actions and events are then functions from state to state. This framework of ideas is used even by many who deny that their formalism contains state variables, and has been deliberately incorporated into several AI programming languages and representational systems. We used it in the toy blocks world earlier. But a slightly broader view condems it.

Consider the following example (which Rod Burstall showed me

many years ago, but I decided to put off until later). Two
people agree to meet again in a week. Then they part, and one
goes to London, while the other flies to San Francisco. They
both lead eventful weeks, each independently of the other, and
duly meet as arranged. In order to describe this using world-
states, we have to say what each of them is at just before and
just after each noteworthy event involving the other, for each
world-state encompasses them both, being a state of the whole
world. But this is clearly silly.

All we need to know about the other persons history is that
at the time of their appointment it is contained in the same
place as the first persons, and this can be established by its
own train of reasoning. When their histories intersect, indeed,
then the interactions between them need to be taken into account
in an adequate description; but not until then.

There are other problems with the "situations" ontology (it
is very hard to give a reasonable account of continuous
processes, for example: see Allen 1983 for some more), but this
alone is enough to indicate that it is not a suitable foundation
for a theory with any breadth.

Events happen in time, but also in space - they have a where
as well as a when. They are four-dimensional spatiotemporal
entities. So are objects, which have a position and shape and
composition at a given time or period, which may differ at other
times, and have timporal as well as spatial boundaries. All of
which suggests that a basic ontological primitive shouls be a
piece of spacetime with natural boundaries, both temporal and
spatial. I will call these things histories. All the spatial

concepts previously introduced can now be seen as instantaneous spatial cross-sections of histories. Thus, a place is a place-history at a time, and an object in a situation is that situations intersection with that objects history. Histories begin and end: the event of putting four blocks together in a square is the beginning of the history of a platform, and the end of that platform is when and where they are separated from one another. Situations themselves, perhaps now better referred to as time-instants, are themselves histories, although of a very special kind, being spatially unbounded and having temporal boundaries defined by the events between which they are fitted.[10] At the other extreme, spatial features which are permanent - notably, permanent places - are histories which are temporally unbounded but spatially restricted. Most objects in the common sense world fit between these extreems. Examples include the inside of a room during a meeting, Lyndon Johnson while he was president (this is an episode in the longer history of the man), Lac Leman (a permanent history) and the trajectory of USAir flight 130 from Washington to Rochester last Wednesday. This last is an example of a history which is more complicated in shape than just the direct algebraic product of a spatial object and a time-period. The projection of a trajectory onto the spatial reference frame is a path (e.g. an air traffic corridor), but the plane was only in a bit of it at each moment: its history slopes in spacetime.

The situations-actions language can be translated uniformly into a language which talks of histories, by replacing

$$R(o1,\ldots,on,s)$$

by

$$R(a(o1,s),\ldots,a(on,s))$$

where a is the function which intersects a history and a time-instant, yielding a purely spatial object. But it is often more natural to describe histories and their relationships in other ways. "Liquids", this volume, employs the histories ontology to describe an aspect of the world which I do not think could possibly be adequately approached using the situations ontology.

There are several kinds of history, and one does not expect that there will be a very rich theory of histories in general. Such as it is, it seems to be concerned with the relationships between histories and their boundaries, a sort of naive geometry of spacetime. Consider for example a stationary object being hit by a moving one and moving itself as a result. There are at least three histories involved in describing this: two successive episodes of the first object and one - that before the collision - of the second object. Call them A1, A2 and B. The temporal boundary between A1 and A2 is a purely spatial entity which itself has a spatial boundary (the surface of the object-at-that-moment: notice that this is the same as the surface-of-the-object at that moment) which is in contact with the (isotemporal) surface of the last moment of B. Something evidently crossed that boundary ("impetus" (McCloskey 1983), probably) and put the first object into a different state: for if nothing had, then there would be no difference between A1 and A2. The event - itself a tiny history - which took place at the point of contact consisted of some kind of transfer between A and

B, and so we must have involved their boundaries, and this is the only place in spacetime where their boundaries intersect.

This vignette of analysis and the "liquids" axiomatisation both illustrate a style of axiomatic description in which histories are classified into types and the kinds of relationship they can have with one another are defined by the nature of their boundary surfaces. Reasoning about the dynamics resembles a process of fitting together a jigsaw of historical pieces in an attempt to fill out spacetime, invoking interface properties of spatial and temporal boundaries at every stage. This appears to be a powerful and general technique, perhaps in part because it adapts so readily to constraint-propagatiobn methods. Forbus (1983) uses a similar idea by partitioning space, as does Allen (1982) by classifying kinds of temporal interval. It depends on the use of taxonomies, i.e. listings of all the possible kinds of history of a certain type (all the kinds of falling history, or all the kinds of timeinterval, or all the ways in which a thing can be supported.[11]

(e) Energy, effort and motion

There seems to be a significant distinction between events which can "just" happen, and those which require some effort or expenditure of energy to keep them going. The difference between falling and being thrown lies almost exactly in this, as far as I can tell. One importance of the distinction lies in the fact that if no effort is expended, then the second kind of history is ruled out, which eliminates a whole class of possibilities from consideration.

This notion of energy is not the physicists one: it is notoriously not conserved, for example (as in hitting ones head against a brick wall, or becoming exhausted by holding a heavy weight). Since real physics has taken the original term away from ordinary language, there are a number of informal terms in use: "oomph", and the German "schwung".

Typically, sources of schwung are of finite capacity and become exhausted in time, although may be self-replenishing. Also typically, schwung can exert force and thereby produce motion (or perhaps one should say rather that it can become motion, and pushing is giving the schwung = force = impetus to the object, c.f. the brief example given earlier).

McClosky, op. cit., has demonstrated convincingly what anyone who has talked to children knows, that naive physics is pre-Galilean. I can still remember the intellectual shock of being taught Newtonian laws of motion at the age of 11. How could something be moving if there were no forces acting on it: but yet, the argument was compelling: for if a surface was completely frictionless then nothing would stop a sliding object. My internal theory had a contradiction at its very center, the realization of which was acutely distressing. Another very convincing intuition is that heavies objects fall faster than light ones,[12] and that released slingshots travel radially outwards.

I believe there are actually two ways of conceptualising motion, which may be analogous to the distinction between large scale space and local metric space: as a displacement or as

a trajectory. A displacement is a change of position, and requires constant effort to maintain: when the effort stops, the motion stops. They are changes of position, having no dynamic or geometric properties. In real-physics terms, they are dominated by friction. Trajectories are the motions of things with impetus. They are smooth motions along paths with a definite shape, and they keep going until they are stopped (when there may well be an impact, in which some or all of the schwung is transferred to other things). Displacement motion is Greek, trajectory motion is Galilean. Concepts such as going, coming, arriving, leaving, to, from, are connected with the former, concepts such as aiming, impact, speed (a quantity space), towards, away from, are connected with the latter. Displacements are really mere transitions from their beginnings to their endings, whereas trajectories have a definite shape, and can be extrapolated in space ant time. Speed is crucial. Walkings are displacements, but runnings have some of the quality of trajectories, and skiings are definitely trajectories. That position changes during the history is true of both kinds of movement, of course: if all we know is that Harry went to the store, it may have been either kind of motion.

(f) Composites and pieces of stuff

Physical objects have many properties and relationships, many of them concerned with external attributes of the object such as shape or position. One category, however, concerns how objects are composed, what they are made of. As far as I can judge, all naive-physical objects are either a single piece of

homogenous stuff, or are made up as a composite out of parts which are themselves objects. The essence of a composite is that its component parts _are_ themselves objects, and that it can (conceptually if not in practice) be taken apart and reassembled, being then the same object. Examples of composites include a car, a cup of coffee, a house, four bricks making a platform. Examples of homogenous objects are a bronze statue, a plank of wood, the Mississippi, a brick. Homogenous objects have no parts, and can only be taken apart by being broken or divided in some way, resulting in _pieces_. Unlike parts, pieces have no independent status as objects in their own right, and the object has no natural internal boundaries which separate them: it comprises a _single_ piece of stuff.

The physical characteristics of a composite depend on those of its parts, but also on the way in which they are arranged. There is a whole collection of concepts which have to do with putting parts together into assemblies: ways of attaching, strength and stability of connections, kinds of relative movement which are possile, how shapes can fit together, adhesive or frictional or lubricated relations between surfaces, etc.: one could put the whole of mechanical engineering in here. Centrally for the theory of composites is that this is _all_ it depends on, so that if a composite is taken apart and reassembled so as to restore all the internal relations exactly, then it will behave in exactly the same way. And it will be the same object. Indeed, parts can be replaced with others - a new engine in a car - and the composite still be considered the same object.[13]   A composite is more than the set of its parts. If we have a kit of

parts for a model airplane, then after assembly all the parts are still there, but the aircraft exists as well as the parts, with its own unique properties. (Notice that the kit then no longer exists. It was also a composite, but of a different kind: not an assembly.)

A homogenous object comprises a single piece of stuff, but is not the same thing as the piece of stuff, since the criteria by which we individuate objects are different from those for pieces. If a statue is melted, the resulting pool is the same piece of (the same) stuff, but a very different object. In fact, the statue is gone forever. Even if the same metal is used in the same mold, the result is a new object. This contrasts sharply with the norm for composites, in which the set of parts is otherwise analogous to the piece of stuff. Pieces of some homogenous objects can be replaced by more of the same stuff and the object retain its identity. This is most obvious for liquid objects such as rivers, but applies also to solid objects, to a more limited extent. If a statue is broken and repaired, its the same statue (compare reassembling a car), although it has invisibly changed, and may now be a composite of the pieces of its former self (contrast reassembling a car).[14] But a piece of stuff is the piece it is, and cannot be added to or subtracted from without becoming a different piece.

Some of the properties of a homogenous object are properties of the object qua object (size, shape), others are properties of the piece of stuff it comprises (amount of stuff:[15] compare number of parts in a composite; color, surface hardness,

rigidity). So long as the object remains the same piece, these both remain unchanged, but when they come apart, some properties can change. Many rivers change color with the seasons: topping up a cup of coffee increases the amount of coffee in the (same) cup: freezing water produces an ice cube.

This last illustrates the distinction between stuffs and physical states (solid, liquid, paste, powder, jelly - a preliminary attempt at a complete list produced over a hundred distinctions, Cunningham 1980). Many stuffs can be put into a different physical state (by heating, cooking, grinding, squeezing, drying, etc.), and much of manufacturing depends on using such transitions to manipulate the object/piece distinction. An example is provided by casting. Take many small pieces of copper and heat them in a crucible. When the copper melts, each piece becomes liquid. Liquids can have no shape, so the copper objects which were the pieces cease to exist. Liquid objects in the same space merge together, so a new, liquid copper object is produced. Now put this stuff into a mold - liquids take the shape of their containers, so the piece of copper now has this shape - and let it cool. Now it is a solid piece of copper and still an exact fit to the mold, so its shape is that of the mold. A new object has been created: an axehead, say. It may have seemed almost like a miracle four thousand years ago.

The parallel distinctions between an object and the piece of stuff which it is, and between a composite and the collection of parts which make it up, make it easy to see why a theory might fail to understand conservation of amount or number during manipulations which change the shape or physical layout of an

object or group: for amount is a property of the piece (or collection), not the object. If that concept is not available, there is no special reason why amount should be preserved, and many examples where it clearly isn't: rivers can get bigger and cause floods, for example. But when the concept is available and is used properly, conservation of amounts is very obvious, since amount of stuff in a piece is a property of the piece: and it is the very same piece after the transformation as before: nothing about it has changed. An ontological shift such as this may provide a convincing amount of the well known phenomenon, first noted by Piaget, of children's sudden acquistion of the "concept" of conservation. Notice however that conservation is not a concept, but a theorem.

(g) Individuation

Establishing criteria for individuation must be done not only for objects but also for spaces, times, histories, quantitites and any other kind of individual in our conceptual universe. When do we ascribe the status of being an individual thing to a piece of the world, since even the purely physical world can be carved up into pieces in arbitralily many ways? I do not think there is a single neat answer, and there need not be: every kind of thing can have its own kind of reason for being a thing. But there do seem to be some general criteria.

We cut up spacetime into pieces so as to (1) keep important interactions as localized as possible: places are pieces of habitable space which are insulated from one another (by distance or by barriers); objects have a complete bounding surface which

separates them from the rest of the world, and (2) to make the interactions as describable as possible. A square of blocks is a platform -, a composite object - if we plan to stand something on it; for in that case we need its top surface to describe the on relation, so we need the object whose surface this is.

Solid objects have a shape (perhaps one that can change within some constraints, like that of an animal) and, while composites can have pieces replaced and retain their integrity, they tend to stay fairly stable. Liquid objects, on the other hand, are defined by their solid containers, and may be in a state of continual overhaul, like a river. The full story is more complex, however, since if the river dries up and refills it is the same river, while if I drink all my coffee, I go to get another cup. (I do not yet understand these distinctions, and would welcome comments.)

The difference between an object and the piece of stuff which it comprises, discussed above, seems to run through many parts of naive physics, and perhaps all of common sense reasoning. The general phenomenon is that one history is an episode of two different histories, each corresponding to a different way of identifying an individual. "Liquids", this volume, describes a particularly intricate example: pouring one glass of water into another.

An important general point is that we do not want anything like universal individualhood. Common sense is prolix - many kinds of entity - but also very conservative - very few entities of each type. This contrasts with more "universal" schemes such as nominalism, in which any piece of spacetime can be an

individual, allowing such things as the sphere of radius 20 meters centered on my left thumbnail _now_, during the month of August 1980 (say). Devotees of higher-order logic as a representational vehicle should realize that when one quantifies over all properties, this similarly means _all_ (describable) properties, such as being further north than the oldest plumber born in Philadelphia. Axiomatic theories must be very careful of comprehension axioms and schemes which guarantee the existence of entities: they should always state the relationship of the new thing to the other things on whose existence it was predicated. Thus we can speak of the space _between_ two walls or _behind_ a door, the falling history which is just _after_ and _beneath_ the moment and place where the object loses its support, and so on. In each case the relations which define the existence of the new entity also attach its boundaries to existing objects.

The use of public global metric coordinate frames restores unrestricted comprehension by the back door, for by using these we can describe the "undescribable" entities: _any_ piece of 3-space, such as an air traffic corridor. The resulting ontological freedom and uniformity may be why coordinate systems are so essential in (real) science.

(h) _A sense of scale_

We seem to be remarkably good at imagining big and small things. One can imagine oneself inside a dolls house, or cupping the galaxy in ones hands. It is as though all our spatial intuitions have a free size parameter, which, while having a normal everyday default setting, can be adjusted so as to bring

other things into their range. The incredible shrinking woman had the misfortune to have her actual size controlled by it. We sophisticated adults know this is impossible, but the idea certainly makes conceptual sense, which it would not if things and spaces had fixed sizes in our conceptual world.

This sliding sizescale seems to be one of the sources of the intuition of continuity in the physical world, and of such geometric abstractions as points and lines. A dot, no matter how small, does have a size (or we wouldn't be able to see it, for example). Imagine it blown up, or equivalently oneself shrunk to match, and it would become an area, a place to be in. Then that space has tiny dots in it, being just like ordinary space. These are invisible in real space, or course, but they are certainly there, for how could it be otherwise? Just turn up the magnification and one would see them. And it must be like that all the way down, since one could always keep on turning the magnification up. That second-level dots are invisible in real life is shown from the observation that real dots are invisible from the next level up, achieved by looking at something from a long way away, so that it becomes small. Since - a basic assumption about scalechange - it doesn't really matter which level one is at, the interlevel relationships must be transparent to shrinking and expansion as well. Mathematical points are now infinitely small dots, which are things that would appear dotlike at all levels. They aren't real physical things, because any real thing has a size and so would eventually stop looking like a dot, but points always resist magnification. (The stars are examples from real life, in fact.)

7. GETTING IT DONE

One objection to the naive physics proposal is that it is
impossibly ambitious: that we don't know enough about
formalisations to embark on such a large representational task;
that it would take centuries, etc. Ultimately the only answer to
such objections is make the attempt and succeed, so all I can do
here is to convey my reasons for feeling optimistic. There are
five.

The first is based on my experiences in tackling the
'liquids' problem, which I had long believed was one of the most
difficult problems in representation theory.   The idea of
quantifying over pieces of space (defined by physical boundaries)
rather than pieces of liquid, enabled the major problems to be
solved quite quickly, to my surprise. The key was finding the
correct way of individuating a liquid object: the criterion by
which one could refer to such a thing. I believe a similar
concern for individuating criteria may well lead to progress in
other clusters as well.

The second reason for optimism is the idea of histories
outlined earlier. I believe that formalisations of the physical
world have been hampered for years by an inadequate ontology for
change and action and that histories begin to provide a way round
this major obstacle.

The third reason is based on the no-programming
methodology already discussed. To put it bluntly: hardly
anybody has tried to build a large, epistemologically adequate

formalisation. We may find that, when we are freed from the necessity to implement performance programs, it is easier than we think.

The fourth reason is that, as the papers in this volume and (Gentner and Stevens 1983) attest, physical intuitions seem to be relatively accessible by such techniques as in-depth interviewing. This was surprising (to me) and encouraging. A common view in AI is that, while expertise is 'surface' knowledge and can be extracted by the expert system builders fairly easily, common sense knowledge is 'deeper', more firmly buried in native machinery, and that to extract it would be much more difficult if not impossible. But it seems not: basic physical intuitions are near the 'surface'. [16]

The fifth reason is that there is an obvious methodology for getting it done, and this methodology has, in recent years, proved very successful in a number of areas.

Within AI, it has come to be called 'knowledge engineering', but essentially the same technique is used by linguists. It works as follows. In consultation with an 'expert' (i.e. a human being whose head contains knowledge: one knows it does because he is able to do the task one is interested in), one builds a preliminary formalisation, based upon his introspective account of what the knowledge in his head is. This formalisation then performs in a particular way, and its performance is compared with that of the expert. Typically it performs rather badly. The expert, observing this performance of the formalisation in detail, is often able to pinpoint more exactly the inadequacies

in his first introspective account, and can offer a more detailed and corrected version. This is formalised, criticised and corrected: and so on. Typically, the expert, continually confronted with the formal consequencies of his introspections, becomes better at detailed introspection as time goes by.

In 'knowledge engineering', the expert is a specialist of some kind, and the formalisation is, typically a collection of condition-action rules which can be run on a suitable interpreter: a very modular program, in a sense. In linguistics, the formalisation is a grammar of some sort which assigns syntactic structures to sentences, and the expert is a native speaker. In both areas, the technique has proven extremely successful.

I believe this process of formalisation, confrontation against intuition, and correction, can also be used to develop naive physics. Here is a domain in which we are all experts, in the required sense. The performance of a formalisation is, here, the pattern of inferences which it supports. Performance is adequate when the 'experts' agree that all and only the immediate, plausible consequences follow from the axioms of the formalisation.[17] It seems to be sound to have several 'experts' involved, as it is easy to miss some obvious distinctions when working alone.

The sheer size of a plausible formalisation should give one pause, however. To even write down ten thousand axioms is not a light task. This can only be a group effort.

The ideal way to make progress is to have a committee. Each member is assigned what seems to be a cluster, and has to try to

formalise it. They tell one another what they require from the other clusters: thus the 'histories' cluster will need some 'shape' concepts, and the 'assemblies' cluster will need some 'histories' concepts, and son on. Fairly frequently, the fragmentary formalisations are put together at a group meeting, criticised by other members (in their common sense 'expert' role), and tested for adequacy. I will anticipate that some clusters will dissolve, and new ones will emerge, during these assembly meetings.

Initially, theformalisations need tobe little more than carefully-worded English sentences. One can make considerable progress on ontological issues, for example, without actually formalising anything, just by being very careful what you say. The "mental modelling" field is at this stage now. But soon it will be necessary to formalise these insights and unify them into the common framework of a broad theory, and this is a new kind of task. It is here that the importance of a common reference language becomes clear, for it is only through this that the minitheories can be related to one another. It seems that this could be a real problem, because everybody has their own favorite notation. Many people find frame-like notations agreeable: others like semantic networks, etc. There is no reason why these, or even more exotic formalisms, should not be used: the only important requirement is that the inferential relationships between the various formalisms should be made explicit. In practice, this means that they should all be translatable into predicate calculus: but this is no problem, since they all are.

All of the suggestions and assumptions I have made are as conservative and minimal as possible. First-order logic is a very simple, basic vanilla flavoured language. Other more structured ideas (procedural representations, frames, p-prims (Disessa 1983), conceptual entities (Greeno 1983), scripts,...) make stronger assumptions about the representational language. It is pessimistic to assume that the a-c graph is connected, and that there is no small collection of primitive concepts. Maybe such special properties of the internal cognitive structure will emerge: but we should discover them, not assume them.

8.   WHY IT NEEDS TO BE DONE.

In the earlier version of this paper I argued at length that tackling a largescale project such as this is essential for longterm progress in artificial intelligence. I will briefly review those arguments here, before turning to other reasons why largescale formalization of "mental models" (Gentner and Stevens 1983) is of basic importance to other parts of cognitive science.

For AI there are three arguments: the importance of scale effects, the need to develop techniques of inference control, and the motivation of adequate representational languages.

AI has the aim of constructing working systems. This might be taken as the defining methodology of the field, in fact, in constrast to cognitive psychology.   But there is a real danger in applying this criterion too early and too rigorously, so that a doctoral thesis must demonstrate a working program in order to be acceptable. Several areas of AI have outgrown this state, but work on knowledge representation is only just beginning to. As I have argued earlier, scale limitations mean that no matter how many short forays into small areas we make, we will never get an adequate formalization of commonsense knowledge. We have to take density seriously, and density requires breadth.

That weak, general techniques of controlling inference are inadequate to cope with the combinatorially explosive search spaces defined by largescale assertional databases is now a matter for the textbooks. The moral is that the inferences-makers need to be informed about what they are doing; they need a theory of control. I will not belabour this point here, but

note that the really large spaces which broad, dense
formalizations yield may need qualitatively different
metatheories of control, or other search processes entirely. I
believe that the study of inferential control (which subsumes
many questions of system architecture generally) is one of the
most important facing AI at present. But until we have some
dense theories to experiment on, we won't know what the real
problems are. Many of the current ideas on controlling deductive
search may be useful only on relatively sparse spaces;
contariwise, richly connected spaces may present new
opportunities for effective strategies (the widespread use of
relaxation, for example, may become newly effective). It would
be interesting to find out, but something like naive physics has
to be done first, otherwise our control theories will be little
more than formalizations of the weak, general heuristics we
already have.[18]

I will bet that there are more representational languages,
systems and formalisms developed by AI workers in the last ten
years than there are theories to express in them. This is partly
because of the pressure to implement already mentioned, but is
also due to a widespread feeling that the real scientific
problems are concerned with how to represent knowledge rather
than with what the knowledge is. When inadequacies arise in
formalizations, the usual response is to attribute the cause less
to the formalization than to a limitation of the language which
was used to express it.[19] Many major recent efforts in the
development of special knowledge representation languages are

concerned with issues which have to do with the structure of the theories which are to be expressed in them. KLONE, for example appears to be a complex notation for describing interrelationships between concepts in a theory, including those between a concept and its constituent parts. The scientific questions of interest are to do with these relationships, not the idiosyncrasies of any particular notation for recording them. But all of this could be carried out in first order logic. The KLONE authors attribute considerable importance to the distinction between the structure of individual concepts on the one hand and the relationships between concepts on the other. In our terms this amounts to an extra layer of structural distinctions added on top of the simple axiomatic theory. Whether or not the distinction is worthwhile, it should not obscure the need to construct the underlying theory itself first.[20]

Progress in building nontrivially large axiomatizations of commonsense knowledge is also of importance to other fields than AI. Any theorising about cognition has to take into account the structure of the internal theories which - if the whole computational view of mind is anything like correct - support it. If this is taken seriously, then large parts of cognitive and developmental psychology and psycholinguistics must refer to internal conceptual structures. This is a truism of cognitive science by now, but what is less widely appreciated is the need to be sensitive to the details of these inner theories. Much work concerns itself with broad hypotheses about the functional architecture of cognitive structure, without paying attention to

the detailed inferences which constitute the internal activities of the system. Some work assumes very simple internal theories, expressed in terms of "schemata", for example, or as an associative network of concept-nodes. But we know that internal theories, if they exist at all, must be extremely large and complex; and we know that we do not yet have any very reliable ideas about their structure, still less about their dynamics. Under these circumstances it seems risky at best to attempt to relate observable behaviour to general hypotheses about cognitive structure. Word meanings in psycholinguistic theorising, for example, often seem to be regarded as atomic entities related by some kind of association. But, as much AI work on language understanding even in restricted domains has shown, words must map into internal concepts in very complex and idiosyncratic ways, and the concepts themselves must be embedded in a network of internal theory, even to make possible such elementary operations as pronoun disambiguation or the interpretation of indirect speech acts.

The medieval alchemists had much empirical knowledge, and very grandiose but simple theories, and some success in relating the two together. Their view of the world attempted to make direct connections between philosophical and religious ideas and the colors and textures of the substances in their retorts. Modern chemistry began when the search for the Philosophers Stone was abandoned for the more modest goal of understanding the details of what was happening in the retorts. Cognitive Science is sometimes reminiscent of alchemy. We should, perhaps, give up

the attempt to make grand, simple theories of the mind, and concentrate instead on the details of what must be in the heads of thinkers discovering it will be a long haul, no doubt, but when we know what it is that people know, we can begin to make realistic theories about how they work. Because they work largely by using this knowledge.

9. IS THIS SCIENCE?

The earlier manifesto ended on a note of exquisite methodological nicety, i.e. whether this activity could really be considered <u>scientific</u>. This second manifesto will end on a different note. Doing this job is necessary, important, difficult and fun. Is it really scientific? Who cares?

10. ACKNOWLEDGEMENTS

It is impossible to name all the people who have contributed
to these ideas. I would, however, like to especially thank Maghi
King, who let me get started; and Jerry Kobbs who made me finish.

FOOTNOTES

1. This touches on a basic terminological ambiguity. Shall we regard an axiom as a statement _in_ a logic; or as a new _rule_ to be _added_ _to_ the logic, so that the logic is somehow made stronger but the axiomatisation is not enlarged? One always has the option: the second route tends to lead to less expressive but operationally more efficient systems, since a rule can often be neatly characterised as an axiom with a restriction imposed on its use, so that less can be inferred from it. I think we should take our axioms unrestricted for a while, until we can see more realistically what sorts of restriction we shall have to impose on their inferential behavior to achieve practical systems.

2. There are two versions, in fact: "all nameable", which you get with the rule or schema, and "all", which can't be enforced by any schema or rule or computational device of _any_ kind, since the set of theorems is then not recursively enumerable. If anyone claims to have implemented a reasoning system which can handle full higher order reasoning, he is wrong.

3. I omitted _clear_ deliberately. It has an explicit definition and could be eliminated entirely at no real cost of expressive power. Having that token makes axioms more compact and deductions shorter, but it does not enable us to

say anything new, since we could have replaced it everywhere else by its definition and gotten an equivalent set of axioms. Definitions don't add to the expressive power of a theory.

4. This is usually presented, in textbooks of elementary logic, in a rather formal, mathematical way: and this fact may have given rise to the curious but widespread delusion that a first-order model is merely another formal description of the world, just like the axiomatisation of which it is a model; and that the Tarskian truth-recursion is a kind of translation from one formal system to another (e.g. Wilks, 1977). This is quite wrong. For a start, the relationship between an axiomatisation and its models (or, dually, between a model and the set of axiomatisations which are true of it) is quite different from a translation. It is many-many rather than one-one, for example. Moreover, it has the algebraic character called a Galois connection, which is to say, roughly, that as the axiomatisation is increased in size (as axioms are added), the collection of models - possible states of affairs - decreases in size. It is quite possible for a large, complex axiomatisation to have small, simple models, and vice versa. In particular, a model can always be gratuitously complex (e.g. contain entities which aren't mentioned at all in the axiomatisation). But the deeper mistake in this way of thinking is to confuse a formal description of a model - found in the textbooks which are a mathematical approach to

the metatheory of logic developing with the actual model.
This is like confusing a mathematical description of Sydney
Harbour Bridge in a textbook of structural engineering with
the actual bridge. A Tarskian model can be a piece of
reality. If I have a blocks-world axiomatisation which has
three block-tokens, 'A', 'B', and 'C' and if I have a
(real, physical) table in front of me, with three (real,
physical) wooden blocks on it, then the set of those three
blocks can be the set of entities of a model of the
axiomatisation (provided, that is, that I can go on to
interpret the relations and functions of the axiomatisation
as physical operations on the wooden blocks, or whatever, in
such a way that the assertions made by the axiomatisation
about the wooden blocks, when so interpreted, are in fact
true). There is nothing in the model theory of first-order
logic which a priori prevents the real world being a model
of an axiom system.

5. In fact, it wouldn't really. To really capture the notion
of 'above', it is probably not enough to stay even within
naive physics: one would have to go into the various
analogies to do with interpersonal status for example.
(Judge's seats are raised: Heaven is high, Hell is low: to
express submission, lower yourself, etc.) Only a very broad
theory can muster the power (via the Galois connection of
model theory) to so constrain the meaning of the token
'above' that it fits to our concept this exactly. (Imagine
a world in which the 'status' analogy was reversed, so that
to be below someone was to be dominant and/or superior to

them. That would be a possilbe model of naive physics, but not of the larger theory of common sense: and it would be a very different world from ours.)

6. Although perhaps one could not fight _too_ hard. It is quite plausible that we might have several minitheories in our heads for some concepts. Perhaps we use one, over-simplified but useful, theory a general utility version and also special-purpose theories to handle idiosyncratic cases (such as porous solids in a theory of liquids). Or, more interestingly, a more sophisticated theory which can handle a very wide range of phenomena but is only invoked when needed (such as an atomic theory to explain porosity, c.f. Lucretius).

7. Perhaps arbitarily or perhaps only seven plus or minus two.

8. It may be felt that this concern with mathematical technicalities is out of place in judging the appropriatenep of an axiomatization, since people don't think about mathematics in everyday affairs. This reaction is inappropriate, however. We are judging the goodness of fit between a formal theory and intuitive reasoning. Intuition seems quite clear on all these matters of touching, which, when formalized, easily yield consequences which are the formal translates of very unintuitive ideas. That the formal derivation uses mathematical ideas is irrelevant to the failure of the match between theory and intuition.

9. And speaking as someone _with_ mathematical training, I think this consists in large part of becoming able to simply

ignore the clash with raw intuition, rather than reconcile
it. A point has position but no extent. How many are there
in a one-inch square, then? Such questions have no answer,
and the training enables one to face this situation with
equinamity. If points really were common-sense dots, there
would <u>have</u> to be an answer.

10. If times passing is represented by a measuring scale, then
we might say that time-instants formed a quantity space with
the measure function defined by a clock. On this account,
the division of the conceptual time into discrete situations
could be seen as the structuring of the past induced by the
clock from the scale. This is how we make appointments to
meet: they depend on there being a public clock and
associated measuring scale.

11. I think there are six. It can be resting on something which
is bearing its weight; hanging from something; attached to
something; floating on liquid; floating in the air - if it
weighs nothing, and then only for a while - or flying, which
takes continual effort.

12. Galileo's own argument why not is beautiful. Consider, he
says, a stone cracked in half, falling alongside an
identical one not split. Let the two halves separate just
slightly. Will the split stone then suddenly decelerate?
Surely not. If so, let the two halves just drift together
and momentarily reunite: will it then accelerate? I tried
this argument out on an intelligent ten year old, but he was
unconvinced, arguing that the two halves would drift apart
vertically, one falling faster than the other, even though

they were identical. Why? - because two things <u>never</u> fell at <u>exactly</u> the same rate. Exasperated by this extraordinary obtuseness, my colleagues and I improvised a demonstration using two pennies. Within the limits of experimental error we could achieve at the dinner table, the child was right.

13. Borderline cases suggest themselves. If one simultaneously replaces everything but the body shell of a car, is it the same car? I think one can say yes, or could alternatively claim that this was a new car: but in that case, the body has been taken <u>from</u> the original car.

14. Primitive atomic theory could be summarized as the idea that homogenous objects are really composites of atoms, and only atoms are truly homogenous (Lucretius). This explains why the recast statue is a new object: the interatomic relationships have changed. If one could get each atom back in the right place, it <u>would</u> be the same statue.

15. Amount is a more basic idea than mass or volume. It takes considerable education to learn to distinguish these.

16. Probably this whole depth metaphor is a mistake, like every other simple metaphor of the mind.

17. In fact, this is a weak notion of adequacy: the stronger notion would be that the derivations of the plausible consequences were also plausible. Attempting to use this stronger notion gives rise to severe methodological problems, since it requires one to have '2nd-order' introspections. Linguistics has an exactly analogous notion of strong adequacy for a grammatical theory, and suffers

exactly similar methodological difficulties.

18. The felt need for a nontrivially complex axiomatisation to try out search heuristics on was my original motivation for embarking on this whole enterprise.

19. This may be connected with the fact that in Computer Science generally, development of programming languages is a respectable academic concern, while the development of particular programs isn't. After all, who knows what a language might be used for, especially a general-purpose language? And knowledge representation systems are almost invariably proud of their generality. This attitude is especially easy to comprehend when the Krep language is considered a species of programming language itself, which was a widespread confusion for several years.

20. The deliberate eschewal of control (= computational) issues in the naive physics proposal represents a very conservative approach to questions of such structuring. First order logic makes very weak assumptions about the structure of theories couched in it, almost the weakest possible. They can be summarised as: the universe consists of individual entities, with relations between them. Nothing is said about the nature of the entities. (An attempt to find an area where this "discreteness" assumptions breaks down was what led me to the liquids formalisation, and individualisation assumption was, unexpectedly, crucial to its success.) It makes no assumptions whatever about control. Any insight into theory

structure which is obtainable within naive physics must be readily transferable to more elaborate notations or systems of representation, therefore. It seems wisest, at this early stage in the development of large "knowledge bases", to be as conservative as possible. One might think that attempting to use firstorder logic as a representational vehicle would be doomed to failure by its expressive inadequacy. In fact, however, the limitations seem to be on our ability to think of things to say in it.

REFERENCES

Allen, J. (1983) Towards a General Theory of Action and Time, _AI Journal_ (to appear).

Cunningham, J. (1980), University of Sussex. personal communication.

Forbus, K. (1981) Qualitative Reasoning about Space and Motion. _TR-615_, MIT AI Laboratory, Cambridge, Mass.

Gentner, D. & Stevens, A. (1983) (eds.) _Mental Models_, Lawrence Erlbaum Associates.

Haak, S. (1973) Do we need "fuzzy logic"? _Ms._ University of Warwick.

Hayes, P. (1977) In defense of logic. _Proc._ _5th_ _IJCAI Conference_, MIT.

Hayes, P. (1978) The Naive Physics Manifesto, (in) _Expert Systems in the Micro-Electronic Age_ (ed. D. Michie), Edinburgh University Press.

Herskovits, A. (1982) _Space and the Prepositions in English: Regularities and Irregularities in a Complex Domain_. Ph.D. thesis, Stanford University.

Howard, I.P. (1978) Recognition and knowledge of the water-level principle. _Perception_, 7, 151-160.

McCarthy, J. (1957) Situations, Actions and Causal Laws. _Memo 1_, Artificial Intelligence Project, Stanford University.

McCarthy, J. & Hayes, P. (1969). Some Philosophical Problems from the standpoint of Artificial Intelligence. _Machine Intelligence 4_ (ed. D. Michie & B. Meltzer), Edinburgh University Press.

McDermott, D. (1977) Artificial Intelligence and Natural Stupidity. Mind Design (ed. J. Haugeland), Bradford Books.

McCloskey, J. (1983) Naive theories of motion, in Gentner & Stevens 1983.

Poston, T. (1972) Fuzzy Geometry. Ph.D. thesis, University of Warwick.

Pylyshyn, Z. (1979) Computational Models and Empirical Constraints. Behavioral and Brain Sciences

Wilks, Y. (1977) Good and Bad arguments about Semantic Primitives. Memo 42, Department of Artificial Intelligence, University of Edinburgh.

Zeeman, C. (1962) The topology of the Brain and Visual Perception. Topology of 3-Manifolds (ed. K. Fort). Prentice-Hall.

Cognitive Science Technical Report List

University of Rochester

1. Steven L. Small and Margery M. Lucas. Word expert parsing: A computer model of sentence comprehension. May, 1983.

2. Gary S. Dell. The representation of serial order in speech: Evidence from repeated phoneme effect in speech errors. May, 1983.

3. Henry E. Kyburg, Jr. Rational Belief. May, 1983.

4. Alan H. Schoenfeld. Beyond the purely cognitive: Belief systems, social cognitions, and metacognitions as driving forces in intellectual performance. May 1983.

5. David A. Taylor, Jung-Oh Kim, and Padmanabhan Sudevan. Representation of linear orders. May, 1983

6. Alan H. Schoenfeld. Episodes and executive decisions in mathematical problem solving. May, 1983.

7. Alan H. Schoenfeld. The wild, wild, wild, wild world of problem solving (a review of sorts). May, 1983

8. Ellen Carni and Lucia A. French. The acquisition of before and after reconsidered: What develops? June, 1983.

9. Susan M. Garnsey and Gary S. Dell. Some neurolinguistic implications of prearticulatory editing in production. July, 1983.

10. Patrick J. Hayes. The second naive physics manifesto. October, 1983.

Note: Requests for U of R Cognitive Science Reports should be addressed to either the author or to Cognitive Science Reports, Department of Philosophy, University of Rochester, Rochester, NY 14627.