Expert Systems in the Micro-electronic Age

EDITED BY DONALD MICHIE

FOR EDINBURGH
UNIVERSITY
PRESS

22 George Square, Edinburgh
ISBN 0 85224 381 2
Printed in Great Britain by
Redwood Burn Limited
Trowbridge, Wiltshire

THE NAIVE PHYSICS MANIFESTO

Patrick J. Hayes

The garden still offered its corners of weed, blackened cabbages, its stones and flowerstalks. And the house its areas of hot and cold, dark holes and talking boards, its districts of terror and blessed sanctuary; together with an infinite range of objects and ornaments that folded, fastened, creaked and sighed, opened and shut.

from 'Cider with Rosie'
by Laurie Lee

0. 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 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.

In what follows I will outline the proposal and distinguish it from some others, superficially similar; discuss some of the general issues which arise; argue that it *needs* to be done; that it *can* be done; and outline a way of *getting* it done. Along the way I will outline the theory of meaning which this proposal assumes, and criticize some others.

1. THE PROPOSAL : SUMMARY

The proposal is to construct a formalisation of a large part of ordinary everyday knowledge of the physical world. 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. The proposal is *not* to develop a new formalism to express this knowledge in. Although we recognise that formalism-hacking may become necessary in due course, we believe that existing, well-understood formalisms have many, as yet unexplored, possibilities. The proposal is *not* to write a program which can solve problems, or plan actions, or whatever, in the formalism. Although it is important to bear control and search issues – in short, *computational issues* – in mind, we propose to deliberately postpone detailed consideration of implementation. All too often, serious work on representational issues in AI has been diverted or totally thwarted by premature concern for computational issues.

The formalism we propose should have the following characteristics (we elaborate on these later):

(i) *Thoroughness.* 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 thoroughness will never be perfect. Nevertheless, we should *try* to fill in all the major holes, or at least identify them.

(ii) *Fidelity*. 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. Again, since the world is infinitely detailed, perfect fidelity is impossible: but we should try to do better than the very low fidelity of the common 'toy problem' axiomatisations, in which, for example, the relationship of one block being 'above' another is merely a partial ordering, so that the integers are a possible model of the axioms.

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

(iv) *Uniformity*. 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.

(As I shall emphasize later, I also think it is methodologically important to allow the use of a variety of formalisms: often a particular subarea can be neatly expressed in some idiosyncratic way. But there is no contradiction: for we will also insist that such idiosyncratic formalisms are systematically reducible to the basic formalism: they will be regarded as 'semantic sugar'. This is important: for although *computational* properties of a representation may depend crucially upon the use of such idiosyncratic formalisms, there must be a common *representational* framework within which the meaning-content of any piece of representation can be related to that of any other.)

I 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 works, like building model railways; and one's students can get PhD'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). Regrettably, the little compromises and simplifications needed in order to get the program to work in a reasonable space or in a reasonable time, can often make the representation even less satisfactory than it might have been.

This is not to say that computational questions should be *ignored* in constructing the proposed formalisation. For example, the question of the *length of derivations* of ordinary common-sense inferences is important, and our notion of 'density' has direct computational consequences, for example for storage and retrieval strategies. But the construction of 'demonstration' programs seems to serve no really useful purpose (McDermott 1977 argues similarly).

I emphasize this point because there is 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. This may be partly to blame for the dearth of really serious efforts in the representational direction, and the proliferation of programs and techniques which work well (or sometimes badly) in trivially small domains, but which are wholly limited by scale factors, and which therefore tell us nothing about thinking about realistically complicated worlds. (Backtracking search and the STRIPS representation of actions by add and delete lists are two good examples. I suspect that production systems are another).

Ideally, one should in principle be able to get a working program from the formalisation by assuming a particular inference mechanism and adding *further* information at the meta-level, which 'controls' the inferences this mechanism performs (Hayes 1973, Kowalski 1977, Pratt 1977, Bundy 1978). Looked at this way, a formalisation can be thought of as a 'core' of inferential abilities, whose appropriate deployment at any moment for a particular task has to be further specified. (However, this pleasant picture has no doubt to be modified to account for idiosyncratic representations which have especially desirable computational properties.)

The decision to postpone details of implementation can be seen as an implicit claim that the representational content of a large formalisation can be separated fairly clearly from the implementational decisions: this is by no means absolutely obvious, although I believe it to be substantially true.

(b) It is *not* 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 it all out in KRL, or semantic networks, or 'fancy' semantic networks of one sort or another, or what have you; well, that's fine. The important point is that one *knows what it means*: that the formalism has a clear *interpretation* (I avoid the word 's*m*nt*cs' deliberately). At the level of interpretation, 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 (Hayes 1977, 1978a).

Secondly, let me emphasize again that idiosyncratic notations may sometimes be useful for idiosyncratic subtheories. For example, in sketching an axiomatic theory of fluids (Hayes 1978b), 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. Other examples are the 'evaluable' predicates and functions sometimes included in theorem-proving preograms, where for example the term '(plus 2 3)' is *evaluated* to the constant '5', no axioms being provided for arithmetic expressions. But this can always * be regarded as a computationally efficient way of representing the same *meaning* as would be represented by the (infinite) collection of axioms '(plus 2 3) = 5', '(plus 2 2) = 4', etc.

Thirdly, first-order logic as it stands is almost certainly not rich enough and will need to be extended. I have already found two extensions which I think are necessary: quotation, so that the formalism can describe its own formulae; and a sort of non-unique Skolem function, similar to Hilbert's ϵ-symbol. The notion of the default may be another (although I have not felt any particular need for this concept so far). I expect that such extensions will arise naturally through difficulties in using the formalism; and as I think it is dangerous to try to predict such difficulties ahead of time, I will not try.

3. THE AXIOM-CONCEPT GRAPH: CLUSTERS AND DENSITY

Let us imagine that a naive physics formalisation exists (I tend to believe that it does, in fact, inside my head), and try to analyse its structure. It consists, fundamentally, of a large number of assertions, involving a large number of (nonlogical) symbols: relation symbols, function and constant symbols (*or*: frame headers, slot names, etc; *or*: node and arc labels, etc. In future I will not bother to re-emphasize these obvious parallels, but will assume the reader is aware of them.) For a neutral word, let us call these symbols *tokens*.

The *meaning* of the tokens is defined by the structure of the formalisation, by the pattern of inferential connections between the assertions. This structure can be very complex, but we can make some fundamental points by treating it in an essentially qualitative way.

Let us say that a formalisation is *dense* if, for each token, there are many axioms involving it. A dense formalisation has many links between the separate concepts expressed by tokens in the formalisation. Density is clearly a matter of degree. Formalisations which are not dense in this way (*sparse* formalisations) are unsatisfactory, since they do not pin down exactly enough the meanings of the tokens they contain. If all a formalisation says about the relation *above* (in the blocks world) is that it is the transitive closure of *on*; and all it says about *on* is that if nothing is *on* a block then you can pick it up; and all it

*This remark slurs over a minefield of technical difficulties. It is not appropriate to go into these here, however.

says about *picking* up is that after *picking*, the block is *held*, etc. (in the familiar way); if this is *all* that it says with these tokens, then one can hardly say that the concepts of *on, above*, etc. are represented in the formalisation at all. For, these concepts have connections to many other concepts as well, in our heads. If one thing is above another, there are all sorts of consequences. Maybe the first will fall on the second, if its supports fail: there are consequences for the relative appearance of the two objects: the top one might provide a shelter for the bottom one: if the bottom one supports the top one, then the bottom one will be under some strain due to the weight of the top one: and so on. We need to try to capture this *richness* of conceptual linking.

(It is worth emphasizing that the view of meaning espoused here differs profoundly from the view which holds that tokéns in a formalisation are essentially words in a natural language (Wilks 1977). According to this latter view, the tokens *do* represent the concept intended, by *fiat*: they are 'semantic primitives', out of which all other meanings are composed. I will return to this idea shortly.)

Any formalisation which hopes to approach the richness of our own conceptual apparatus must be dense. Of course, density is not a *sufficient* condition for success: it is not difficult to invent wholly useless axiomatisations of arbitrarily high density.

It is useful to consider a simplified model of a formalisation. Imagine a graph whose nodes are tokens of the formalisation, and whose arcs correspond to axioms: an arc links two nodes if the corresponding axiom contains those two tokens. (Strictly speaking, this has to be a polygraph (Landin 1970), since axioms may well contain more than two tokens. However, we will be making only heuristic use of the idea, in any case, so technicalities are inappropriate.) We will call this the axiom-concept (a-c) graph. The formalisation is dense if the a-c graph is highly connected, sparse if the graph is sparse. However, we cannot expect density to be uniform: there will be more dense clusters of concepts with many relationships between them, less tightly linked to the rest of the formalisation.

Identifying these 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'. 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.

There is also a useful notion of a super-cluster: a cluster which is related to a large number of other clusters. I think that the collection of concepts to do with three-dimensional shape and orientation are a super-cluster in our own mental conceptual structures: concepts such as above, below, tall, fat, wide, behind, touching, resting on, angle of slope, edge (of a surface), surface (of a volume), side, vertical, top, bottom, . . . These obviously have many internal relationships: they form a cluster. They also must appear significantly in whatever conceptual frameworks underlie visual perception and locomotion in space; they are crucial in describing assemblies; in the theory of liquids (Hayes 1978b); also, I believe, in the descriptions of physical actions and events (Hayes 1978c); and so on.

Superclusters can be recognised by the fact that they crop up, in this way, in a variety of other clusters. Other plausible candidates for superclusters include a theory of measuring scales (which should provide such notions as accuracy, vagueness, utility for various kinds of tasks), a theory of time-measurement, and the collection of notions concerned with *inside, outside, containment* and *ways through* from one place to another.

This suggestive terminology of clustering should not be taken literally: I do not mean to suggest that there are many sharply isolated parts of our conceptual structure which can be developed in total isolation from all others. And the 'a-c graph' model of an axiom system is itself over-simplified in several gross respects, in any case. Nevertheless, I think the basic idea, that there are collections of concepts which have close connections between one another, is substantially correct and quite important.

4. THE A/C RATIO AND REDUCTIONIST FORMALISATIONS

Let me turn now to an even cruder model of a formalisation: the *ratio* of axioms to concepts (the *a/c* ratio). For a dense axiomatisation, *a/c* will be large. Any interesting axiomatisation will have *a/c* greater than one; but there *are* interesting axiomatisations in which *a/c* will be very close to unity.

Consider an axiomatic set theory. The idea, for foundational research, is to have a *small* axiomatic theory (e.g. Zermelo-Fraenkel set theory which has $c = 2$, viz. 'ϵ' and '*set*', and $a = 8$, so $a/c = 4$) which enables one to *define* a large number of mathematical concepts (e.g. the integers can be defined as sets of any one of several special kinds; the rationals are sets of pairs of integers; the reals are sets of infinite sets of rationals . . .) in such a way that the desired properties of these concepts (e.g. the principle of induction for integers: the continuity of the real line) *follow from* the structure of these definitions, and the axioms of the basic theory. It is important to realise that these properties of the defined concepts are *theorems* of the axiomatisation which consists of set theory together with the definitions of the concepts. They are not axiomatic assumptions themselves, needed to pin down the meaning of the introduced concepts: the definitions pin down the concept as completely as it can be: all else then follows. Mathematics is reduced to a series of *lemmas* to set theory: or at least, that is the idea. (When one thinks of it this way, it seems

almost incredible that such an audacious programme should have so nearly succeeded.)

I want to emphasize how different this approach to capturing meaning in a formalism is, from the axiomatic approach to naive physics which I am proposing. Set theory is reductionist in the extreme: I am urging a richly connected formalisation, with many interactions between assumptions. The reductionist approach leads, indeed, to axiomatic theories, but they are extraordinarily *sparse*.

Consider the effect of adding a definition of a new concept *to a formalisation*. This increases both a and c by one. If a/c is large, this will decrease it appreciably. (Indeed, one will have to add approximately (a/c) many axioms to bring the ratio back to what it was.

This emphasizes what is intuitively clear, that definitions which do not pay their keep by introducing concepts which are going to be of some general use, are probably a mistake: they dilute the formalisation.) Suppose however that a and c are both small, say $a = 8$, $c = 2$. Then adding one definition reduces a/c from 4 to 3. Adding another reduces it to 2.5: adding 1000 definitions reduces a/c to 1.059. Clearly, as the number of definitions in the formalisation increases, a/c tends asymptotically to unity. The a-c graph of such a formalisation has one very small cluster at the centre, surrounded by a cloud of nodes each linked radially to a few nodes closer to the centre. It is almost as sparse a connected graph as one could have, containing the concept tokens that it does. It has quite a different 'shape' from the connected, clustered graph of a dense axiomatic theory.

The existence of such a reductionist theory for mathematics is a remarkable fact, and it would indeed be astonishing if one could find an analogous reductionist theory for common sense reasoning: a small-ish collection of concepts, and axioms connecting them, such that all other concepts (e.g. all those expressed by English words) could be defined in terms of these few. In fact, this would be *so* astonishing that I feel confident in asserting that no such small theory exists. And yet, many approaches to the formal representation of meaning, in the AI literature, make such an assumption. These are the 'semantic primitive' approaches, exemplified by the work of Wilks (1975, for example) and Schank (1975, for example). Here, members of a smallish collection of tokens (Wilks 90 or so, Schank 14+n for some n, as yet unknown) are taken as *primitive*. The *meaning* of an English word is then some formal expression built out of these primitive tokens, using some formal apparatus (which in Schank's case is usually presented graphically, although this is not essential). In our terms, the formalisation consists mostly of definitions: its a/c ratio tends to unity, just like axiomatic set theory. In the work of Schank and his students one can clearly see that the axiomatic structure of the 'core' theory – cf. Rieger's 'inference molecules' associated with the 14 primitive action-tokens – is intended to serve the same sort of central organising role that the set axioms play in the development of set theory. That is, desired properties of non-primitive concepts (such as *buying* or *giving*) follow from their definitions, and the meaning

given to the primitives by the core theory. In Wilks' work, there does not seem to *be* any core formalisation at all: we are merely present (cf Wilks 1977) with a list of tokens and a brief description of the concept they are supposed to mean. To think that a *formal* symbol has any meaning other than that specified by the structure of the formalisation in which it occurs, that it has any *intrinsic* meaning, is to make a particularly unfortunate mistake. Wilks, however, takes the view that his semantic primitives are, in fact, *words*, like English words, not mere formal tokens, thus neatly escaping this objection, and explaining why it is not necessary to give any formalisation in which they occur. I confess to still being puzzled as to how it is that his *program* knows what they mean.

This reductionist, semantic-primitives approach to meaning is essentially bound to low-fidelity, low-density representations. Such representations have their uses – they may be adequate for information-retrieval or machine translation applications, for example – and when they *do* work, have some very useful computational properties. But at some point we will have to face up to the problem of representing *detailed* knowledge of the world. This will require the abandoning of the 'definition' view of meaning of tokens. As Wilks says, 'No representation in primitives could be expected to distinguish by its structure the senses of *hammer, mallet* and *axe* . . .'. Perhaps not: but naive physics should be able to.

5. MEANINGS, MODEL THEORY AND FIDELITY

It might be asked: if the meanings of tokens are not specified by definitions, how *are* they specified? In one sense there is no answer to such a question. One cannot point to a particular structure and say, *that* is the meaning of a token. One can only say that a token *means* a concept to the extent that *the formalisation taken as a whole* enables a sufficient number of inferences to be made whose conclusions contain the token, i.e. which mention the concept. This operational definition of meaning can, if the formalisation has an adequate model theory, be recast in an extensional way: a token means a concept if, in every possible model of *the formalisation taken as a whole*, that token denotes an entity which one would agree was a satisfactory instantiation of the concept in the possible state of affairs represented by the model.

Now, to do this requires that it be possible to think of a model as a 'state of affairs'. And since we want to formalise the common-sense world of physical reality, this means, for us, that a model of the formalisation must be recognisable as a facsimile of physical reality: one in which the concept we are interested in can be recognised.

A model for a first-order axiomatisation is a set – the set of entities which exist in the 'state of affairs' represented by the model – and a particular mapping from the tokens of the axiomatisation into this set and the sets of relations and functions, of appropriate arity, over it. 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 axio-

matisation of which it is a model; and that the Tarskian truth-recursion is a kind of translation from the latter to the former: a 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 developing a mathematical approach to the metatheory of logic – 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 actually *be* a piece of reality. If I have a blocks-world axiomatisation which refers to three blocks, called 'A', 'B', and 'C' (i.e. these are the tokens used in the axiomatisation to mention the blocks), 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 Tarskian model theory of first-order logic which *a priori* prevents the real world being a model of an axiom system.

On the other hand, it is also true that many axiomatisations have models which do not contain solid physical objects, but in which tokens denote, say, integers or other symbols. In fact, any first-order axiomatisation which has any model at all (i.e. which is consistent) also has a model in which only symbols exist – this is the 'Herbrand interpretation' in which we let tokens denote themselves. Thus, Tarskian model theory does not guarantee that axiomatisations are 'about' any particular world. It is always possible to consistently believe that the only things which exist are the symbols of the formalisation itself.* This might be called the 'solipsist' interpretation: denying the existence of the external world, while at the same time having an elaborate theory of it.

Given then, that any axiomatisation will have several (usually infinitely many) models, we cannot justify an axiomatisation as an adequate description by merely exhibiting a model of it which is somehow similar to reality. For it may have a very much simpler model than that, and if it has such a simpler

*Although in the usual formalisation of first-order logic, this belief cannot be expressed: and in at least one extension in which it *can* be expressed, its negation – that some things exist which are not symbols – can also be expressed.

model, then the tokens occurring in it mean no more than they mean in that simpler model. This is exactly what I mean by 'fidelity'. A low-fidelity formalisation of, say, the blocks world will admit models which are very much simpler than the intended one: a model, for example, in which the 'blocks' are integers and *above* means greater than, etc.; or perhaps, a model in which 'blocks' are points in a discrete 2-dimensional space, or whatever. An adequate formalisation of a blocks world – a high-fidelity formalisation – will be such that any model of it must have an essentially three-dimensional structure. Thus, SHRDLU's blocks-world axiomatisation (Winograd 1972) uses three-dimensional Cartesian coordinates. (It would be interesting to find a useful but less quantitative way of describing three-dimensional structure: for example, that a rigid attachment needs *three* points of contact: two-legged stools fall over.)

A good guide to the fidelity of a formalisation is how closely its *simplest* model resembles the *intended* model. This is, I think, an excellent argument for a representational language's having a model theory: it gives us a way of testing the fidelity of a representation. It is perilously easy to *think* that one's formalisation has captured a concept (because one has used a convincing-sounding token to stand for it, for example), when in fact, for all the formalisation knows, the token might denote something altogether more elementary, in a very simple model.

(This criterion suggests that we should pay attention to features of representational languages which can be used in a formalisation to insist upon the complexity of the simplest model. In the case of first-order logic, these include functions (the use of which claims that there is a value of the function, applied to any suitable argument), explicit existential assertions (especially 'comprehension axioms' of various kinds – more on this below), the use of equality, and the use of a highly sorted logic (or, in terms more familiar to AI, the presence of an *isa* hierarchy related to the quantifiers: although it can be much more complex than a simple hierarchy, cf. the sorted logic used in Hayes 1971. We might expect, therefore, that these features will be heavily used in developing a naive physics.)

On this account of the meaning of a token, it depends upon the *entire* formalisation of which the token is part. Thus, an alteration to any part of the formalisation can, in principle, change the meaning of every other part of it. And, I think, this is essentially correct: what it means, in introspective terms, is that learning a new fact or acquiring a new concept, is liable to have far-reaching consequences for the ways in which one understands the meanings of other concepts. It also means that people with different formalisations in their heads may understand the same token in different ways. What I mean by 'water' may not be exactly what you mean by 'water': it may be possible to find a substance and a set of circumstances such that I would call it water and you would not (for example, if you had never seen opaque water, you might deny that this opaque liquid which I have in a glass on my desk, was water). And yet, *we might both be right*, since our theories of 'water' may not be identical. And this may even be possible when our beliefs about water (in the direct sense of:

all the assertions which actually contain the token 'water') are identical. Each of us might agree to everything the other said about water, and yet our concepts might be subtly different. The difference may lie in some related concept (such as viscosity, or drinkability) which we understand differently. It may not even be possible to say exactly which tokens we differ on: just that we have somewhat different theories of them. It follows from this that there is no *single* 'meaning' of a token: or at least, that to assume that there is, is to assume that people's cognitive formalisations are identical in structure. Much confusion follows if this is not borne clearly in mind. For example, Wilks has argued (Anderson et al. 1972) that since there are people who have never seen ice, the fact that water freezes *cannot* be part of the meaning of 'water': for if it were, one would have to say that those people did not understand the meaning of 'water': and that is ridiculous. One can cut through this tortuous piece of reasoning by observing that the word evidently means more to some people than to others: and to those who *do* know about ice, the fact that ice is frozen water can well be part of the meaning of 'water'.

However, in order that communication be possible at all, it should obviously be that people's cognitive structures are *similar*: and, as a working hypothesis, we will make such an assumption in developing naive physics. One of the good reasons for choosing naive *physics* to tackle first is that there seems to be a greater measure of interpersonal agreement here than in many fields.

There seems to be a notion of 'distance' in a formalisation, such that the effect of an alteration on the meaning of a token is less, the further away the token is from the alteration. It is not clear to me whether or how this suggestive intuition can be made to stand up. It is tempting to identify this distance with shortest-path distance in the axiom-concept graph, and although this is not really adequate since it ignores the structure of the axioms, it is the best I can do at present.

Thanks to this distance-dilution effect, it seems a reasonable strategy 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.

What I am claiming here is, at bottom, that although the 'definitions' view of meaning is wrong, we can – indeed, must – act *as though* it were correct, in order to make progress. It is good methodological scaffolding.

One last point on the model-theoretic view of meaning. As I have said, any consistent first-order axiomatisation has a model in which there are only

symbols. For all that the simplest such model may be very complex, one might feel that if all it contains are *symbols*, it hardly can be said to be like the real physical world, even if it is in some sense similar in 'abstract' structure.

To answer this objection, we have to talk about the body and sensory input. Imagine the naive-physics formalisation has a (physical) body with sense-organs. The way in which a formalisation can be attached to the physical world is by taking a 'realist' view of the data supplied to it by its perception. Thus, part of naive physics should be a theory of *appearance*: such a theory is now being developed by work in visual perception (especially the work of Marr and Horn, which consciously attempts to relate appearance to physical structure). While it may be that a detailed, hi-fi theory of wooden blocks could all be a dream: block-tokens might denote (say) integers in some models, for example: this would not be true if the theory also specified that if (to oversimplify) one directed one's gaze at a block with *such* a kind of surface, and *such* an orientation and in *such* a kind of lighting conditions, then one would see *such* a kind of image. For, an integer, or a symbol, is not the sort of thing one can direct one's gaze towards. (And, even if it were, in some outré sense, it certainly wouldn't look like a brick.) It might be objected: but you are *assuming* that the tokens for 'directing one's gaze' must denote the physical action of so doing, which begs the question: for there may be an interpretation of these tokens in a model of the integers, say, which also satisfies the axiomatisation. And yes, I am exactly begging the question: I assume that 'motor tokens' – symbols which describe bodily movements – are *directly* related to the body. They constitute a *body image** which has a very special relationship to the (actual) body (I imagine it to be similar to that between a graphic data structure and the physical picture on the screen).

What this assumption means, then, is that a naive physics can be 'connected' to the real physical world because it has a physical body, equipped with sense organs: the notion of *directing* one's gaze (or of *feeling*, or *pushing*, etc.) is *essentially* physical, and has this character because it has a fixed interpretation in the body's sensori-motor system. It is this fixed, physical interpretation of some of the tokens in an axiomatisation which attaches the axiomatisation as a whole must contain real, physical entities and relationships. In some sense, therefore, one would expect that most of the naive physics was 'close' to the visual-appearance (or touching-and-feeling or smelling or hearing) cluster(s) of concepts: that no part of naive physics is *very* remote from sensory evidence. And indeed, when elaborate theories are evolved in real physics, containing concepts which *are* inferentially remote from the evidence, there is usually a general feeling of mild disquiet. One hears talk of *theoretical entities*, for example, (electrons are a good example) in the philosophy-of-science literature. What these discussions seem to me often to lack is a strong enough sense of the way in which even everyday, mundane ideas such as a *piece of wood* or *being wet* are constructs just as theoretical, albeit in a different and more naive theory of the world.

A moral of this for naive physics is that we should be always ready to seize

*I am indebted to Sylvia Weir for introducing me to this notion.

on a chance to relate concepts to sensory or sensori-motor concepts. In working with liquids, for example, I found a notion of movement-in-space very useful: and it has obvious utility in other areas as well. This can be related directly to a visual-perception theory. If you look at a space in which there is movement, then the movement can be *seen*. Ullman (1977) explains in delightful detail how to see it. Again, much of the richness of texture of our introspective common-sense world comes, I think, from our knowledge of *what it feels like when we do things* like pushing, pulling, lifting: from, ultimately, the proprioceptive sensors in our body joints and muscles. I am *not* optimistic that we can capture this richness in a formalisation in the foreseeable future. (It requires the construction of a suitable physical body, equipped with the necessary senses.) But I think we *can* annotate the formalisation by noting the concepts which would be 'attached' to bodily-movement-concepts – and that this would be a useful and interesting exercise.

This whole area of psychosomatic relationships is one which deserves deeper study, and which I believe AI concepts can do much to clarify.

6. THOROUGHNESS AND CLOSURE

One way to have a high a/c ratio, it might seem, would be to keep c small, and to say a lot about a few concepts. And if this could be done, it would indeed be very useful and encouraging: if we could find some small, self-contained groups of concepts which could be formalised in total isolation to a reasonable degree of fidelity.

But there don't seem to be many of these. (Geometrical shape may be one.) The typical situation one finds is that, having chosen one's concepts to start on, one quickly needs to introduce tokens for others one had not contemplated: and in order to pin down *their* meanings, yet more concepts need to be introduced, and the proliferation of tokens seems to be getting out of hand. If one thinks of this as exploring the a-c graph of our conceptual structures, this phenomenon is of course, hardly surprising (especially if we assume, as we must, that the graph is very dense). One needs a sense of direction, to stay within the current cluster while recognising paths into others. But even with such a sense (which will only be developed with experience), the proliferation of necessary concepts seems almost frightening at first.

But this proliferation *must* slow down eventually: for the formalisation is finite. The point of the 'thoroughness' requirement is to *go on until it does slow down*: until one finds that the collection of concepts has closed upon itself. so that all the things one wants to say in the formalisation can be said using the tokens which have already been introduced. In the graphical analogy, until we have *spanned* the entire graph, and need only to add new arcs, filling out the graph until its density is sufficient to capture the meanings of its tokens.

This idea of closure is familiar to anyone who has built toy-world axiomatisations. One finds, suddenly, that there are enough concepts around so that one can say 'enough' about them all: enough, that is to enable the inferences that one had had in mind all along to be made. Closure can be achieved in

very small formalisations: but if a formalisation is closed *and* has high fidelity (so, high density), then it must, I believe, also be thorough: its scope must cover *all* the major concepts of common-sense reasoning. This amounts to claiming that the a-c graph is fairly strongly connected: there are no really isolated subgraphs.

This correlation between thoroughness and fidelity is a matter of degree. To achieve greater fidelity, one will need greater thoroughness. 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 possible model of naive physics, but not of the larger theory of common sense: and it would be a very different world from ours.) A formalisation cannot be deep without being broad, and must be deep to be dense: so a dense formalisation must be deep and broad.

Clusters are exactly partial closures in this sense. A cluster contains a group of concepts which close in on one another to *some* extent: one does need other concepts, but within the cluster there are a long of things to be said about the cluster's own concepts. Clustering is also, therefore, a matter of degree, and depends upon the fidelity, the level of detail.

The whole programme of tackling naive physics in isolation from other parts of common-sense is based on the view that there is a level of detail at which naive physics forms a reasonably close cluster in the larger conceptual structure, and that this is a rich but tractable level of detail: it represents, I believe, an order of magnitude more thoroughness than has yet been achieved; but not, say, ten orders of magnitude.

7. SOME LIKELY CLUSTERS AND THEIR CONCEPTS

In this section I sketch some concrete ideas for concept-clusters. These are only sketches, and vague ones at that: fuller accounts will appear elsewhere, eventually. I do not want to suggest that this is in any way an exhaustive list.

It is quite likely that many of these are not clusters in any real sense. They may, for example, split into pieces on closer investigation; or, new links may be revealed which blur the edges of the clusters. Nevertheless, they seem to me to be quite good places to start exploring.

7.1 Measuring Scales

We need to be able to express *quantities* such as size, extent, amount (of a liquid or powder), weight, viscosity, etc. These seem to be properties of things. But we can measure weight, for example, in pounds or in kilograms, so we have to introduce the notion of various measuring scales which measure the

same physical quantity. We could have various functions from objects to (say) the rationals, called weight-in-lbs and weight-in-kilos, etc.: but this is awkward, unnatural and does not support a very dense collection of axioms. I think we should introduce a notion of an 'abstract space' of *weights (sizes, amounts)*, so that *weight* is a function from objects to weights, and *lbs* (etc.) are functions from rationals to weights, and we can write:

weight (Fred) = pounds (150.32) = kilos (68.25)

These *measure spaces* of weights, sizes, etc. have, I think, a theory of their own. They probably have the structure of a tolerance space (Zeeman 1962), i.e. they have a finite 'grain'. They are notions of approximation, nearness, 'typical' measures of various kinds (normal-sized for an elephant), of inequalities, and of other related matters: and I guess much of this is independent of the particular quantity being measured.

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.

7.2 Shape, Orientation and Dimension

Physical 3-dimensional shape. This cluster is not much investigated, it seems to me, in spite of the considerable work in robot manipulator languages (cf. Bolles 1976). It is also related to topics in visual perception, and as there is more work here, this may be a good way into it. I wish there were more I could say about shape, but I can make only a few loose remarks.

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 when it is against a rigid surface (such as a wall) than when it is free-standing (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 of shape 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.

7.3 Inside and Outside

Consider the following collection of concepts: (inside), (outside), (door, portal,

window, gate, way in, way out), (wall, boundary, container), (obstacle, barrier), (way past, way through).

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 into *pieces* which have physical boundaries, and the ways in which these pieces of space can be connected to one another; and how objects, people 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. 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'.

It is interesting to contrast these ideas with ideas of shape. Here, we are concerned with space as a place to be in: *room to move*, as it were; whereas in describing shape, space is *space occupied* by a substance. There are many concepts useful in both areas, however.

7.4 Histories: Describing Happenings

The now-classical approach to describing actions and change, pioneered by John McCarthy, is to use the concept of *state* or *situation*. This is thought of as a snapshot of the world at a given moment: actions and events are then functions from states to states. This framework of ideas is used even by many who deny that their formalism contains state variables, and has been consciously incorporated into several AI programming languages. I now think however, that it is a mistake or at least a gross over-simplification.

Consider the following example (which Rod Burstall showed me many years ago, but which I did not appreciate at the time). Two people in New York agree to meet a week later in London. Then they go their separate ways: one to Edinburgh, one to San Francisco. Each of them leads an eventful week, independently of the other, and they duly meet as arranged. In order to describe this using situations, we have to say what happens to each of them after every event that happens to the other: for each situation, being conceptually a state of the whole world, encompasses them both.

What we need is a notion of a state of an event which has a restricted spatial extent. By a *history* I mean such an object, viz. a connected piece of space-time, typically bounded on all four dimensions, in which 'something happens' (where I intend this to include the special case of nothing happening).

A three-dimensional spatial cross-section of a history is a place at a certain moment, i.e. a state of that place. Places can be larger or smaller, and can be nested inside one another: a room, a hotel, a street (understood as being the space inside all the buildings which open onto the street) and a city can all be places. A typical history will be, for example, the inside of a certain room from 1.00 p.m. until 4 p.m. on a certain afternoon. Space is, conceptually, made up of places, and space-time is made up of histories, fitted together in a jigsaw pattern. One can also think of a history as *the extension of* (an occurrence of) a process.

Any well-defined object or piece of space can be trivially extended into a history by multiplying it (in the algebraic direct-product sense) by a time-interval, but there are also somewhat less simple histories, such as trajectories, which 'slope' in space-time.

It is very useful to be able to refer to the *shape* of a history as well as of an object or piece of space. For example, a column of falling water (pouring from a jug, for example) defines a history, the shape of which is a vertical cylinder. The top and bottom of this cylinder are of some importance in relating this history to others in which liquid is moving.

Histories can be related to one another in various ways. There are *adjacency* relationships, both spatial (e.g. vertically-above and touching, as in a column of water falling onto a table-top) and temporal (e.g. immediately-following, as in touching a switch, thus starting a motor), and hybrid (as in a collision between aircraft, which is the intersection of two trajectories). There are shape properties and relative-position relationships between places and hence between histories. There are the spatial containment relationships similarly inherited from places and objects, and also temporal containment relationships ('while'). And there are relationships between histories and various global coordinate systems, both of space and of time: what we might call the *address* of a history. There are many possible addressing systems, not all being metric coordinate frames, e.g. room-numbering systems within a building. All of these define what might be called a naive geometry of space-time.

Not every patch of space counts as a well-defined place, and not every patch of space-time as a history. There have to be 'natural' boundaries defining the edges. What counts as a natural boundary is (deliberately) open-ended, but physical barriers are obvious examples, such as the walls of a room.

Since places can be nested (indeed, perhaps every place is *inside* some other place), every event is contained in many (perhaps infinitely many) histories. But for each type of event, there is a smallest which strictly contains it, viz. the smallest one which is spatially bounded by barriers which are opaque to the causal consequences of that sort of event. (*Roughly*, this place is the natural answer to the question 'where did (the event) happen?' (possible answers – in front of the desk; in the living room; in that house: in London)). The importance of this idea has already been mentioned: such barriers limit the extent to which causal consequences of events have to be pursued, and hence make prediction easier. One can predict, from a *static* description of the barrier-geometry, that various kinds of event can affect only a few histories. For example, only a

restricted class of very unusual events taking place in a closed room (large explosions, flooding, large fires), can directly affect histories outside the room.

There are several other important increases in expressive and predictive power which histories give us, compared to the classical situations/actions ontology, but it would take too long to go into more detail here. A fuller account is in preparation.

7.5 Energy and Effort

In making predictions, there is a distinction which seems crucial between events which can 'just happen' (such as fallings) and events which require some effort or expenditure of energy (such as rocks flying through the air). The point being, of course, that if there is no effort being made in a given history, then the latter kind of event is ruled out.

Such a distinction runs counter to the law of conservation of energy, and I think quite correctly so for naive physics (or we could say merely that the intuitive concept of 'effort' does not exactly correspond to the physics notion of 'work'). There are many everyday situations where energy is expended with little apparent result. (Hitting a nail into a brick, for example).

I am not sure what else can be said about the concept of effort: perhaps that sources of effort have a finite capacity (they wear out or get tired). It may be that this is not so much a cluster as a concept which loosely links together a number of other clusters.

Agents can be sources of energy, but the two concepts are distinct, since some actions require no energy (speaking, for example), and there are sources of energy which have no volition. However, it may be that the two are equivalent notions within naive *physics*, and can only be distinguished by the use of 'psychological' concepts such as volition.

7.6 Assemblies

Many solid physical objects are made from parts assembled together in some way: while others are simply a piece of (some kind of) stuff, such as a block of wood. There are a number of concepts which are connected with this idea of an assembly: notions such as being a component of, being a part of (for example, one's hand is a part of one, but not a component: one's liver is (arguably) a component: it is, as it were, detachable), attachment point, assembling and taking apart; glueing, nailing, screwing, etc. . . There are also notions to do with the ways in which assembled parts can move relative to one another (shafts, pulleys, keyways, hinges), and these connect to the spatial-geometry concepts of shape and movement. And there are concepts to do with the mechanical properties of different kinds of stuff: rigidity, hardness, flexibility, being able to cut easily, being able to be glued, etc.

7.7 Support

Objects (or liquids) fall, if left to themselves. To stop them falling they must be supported. I think we can make a short list of all the ways in which a thing can

be supported, as follows.

(a) Something is underneath it, holding it up. Of course, this must be supported too, and so on: but the *ground* does not require support: *it* is the bottom of all support relationships (from which it follows that the ground is not an object).

(b) Something is above it, and it is *hanging from* that thing.

(c) Something is alongside it, and it is *attached to* that thing (cf. assemblies).

(d) It is *floating* on some contained liquid.

(e) It is *flying*, i.e. holding itself up in some way, without touching anything solid. This requires great effort on the part of the flying object, so that inanimate, 'passive' objects cannot fly (although kites may be an exception).

Of these, (a) is by far the safest: in all the others, a failure of some component (respectively: a breaking (e.g. of string), a detaching, a leak, a cessation of effort: all of these are histories of one kind or another) which can mean the ending of the support and hence the sudden beginning of a falling history, and falling histories have dangerous endings (typically). Hence, there is a mini-cluster of concepts around the idea of 'support-from-below': notions such as pile, tower, wall, etc.: stability and the ways it can fail (falling over, crumbling, coming apart, sliding, subsidence).

7.8 Substances and Physical States

There are different kinds of *stuff*: iron, water, wood, meat, stone, sand, etc. And these exist in different kinds of *physical state*: solid, liquid, powder, paste, jelly (jello for U.S. readers), slime, paper-like, etc. Each kind of stuff has a *usual* state: iron is solid, water is liquid, sand is powder, etc., but this can sometimes be changed. For example, many stuffs will melt if you make them hot enough (which for some things is *very very* hot, i.e. *in practice* they can't be melted, e.g. sand; and others will *burn* when heated, e.g. wood or flour). Any liquid will freeze if you make it cold enough. Any solid *can* be powdered if you pulverise it with enough effort and determination, etc. There is no obvious standard way of changing a powder into a solid (but wetting it to get a paste, then drying the paste carefully, sometimes works).

Sometimes we have a separate concept for the same substance in two different states. Sand and rock are a good example. This is, I think, worthwhile when it is (a) extremely difficult to convert one to another, and (b) both occur in nature. (Contrast iron filings, which satisfies (a) but not (b).)

Some substances, left to themselves, *decompose*, i.e. change slowly into some other (useless) substance; or *mature*, i.e. change slowly into some other (useful) substance. Rusting and wet rot are examples of decomposition, cheese-making an example of maturation.

Every physical object which is not an assembly must be made of some stuff, and many of the properties of the object are in fact properties of the stuff of which it is made (rigidity, colour – unless it has been painted – hardness, etc.). It is, I think, important to separate these properties of a thing from those which

are are essentially connected with the shape or structure of the thing. Some objects are essentially defined by one kind of property (a lump of lead), others by other (a building block). Properties such as *weight*, which involve both size and material, have different implications depending upon the kind of object: a heavy lump of lead is a big lump of lead, whereas a heavy building brick must be made of some especially dense material. Solid objects must be made of materials whose physical state is solid, for only solids can be said to have a shape. It follows that if a solid object is heated to the melting-point of the material of which it is made, it must cease to exist as an object, since the requisite state of its substance no longer obtains. This is, I think, a very convincing account of melting.

Cookery would seem to be a good area in which to explore the possible transitions between physical states of various semisolid substances, and manufacturing processes (moulding, casting, forging) another. I think the various methods of measuring amount and quantity might also be a useful area to explore. Wood and metal, for example, are wholesaled by weight or volume (basically), but retailed in various different systems depending upon the shape of the pieces (strip-like or surface-like or solid).

7.9 Forces and Movement

Naive physics is pre-Galilean. I can still vividly remember the intellectual shock of being taught Newtonian 'laws of motion' at the age of 11: how could something be moving if there was no force acting on it? It is interesting to read Galileo's "Dialogue Concerning the Principal Systems of the World" (1632), where he argues very convincingly, from everyday experiences, that Newton's first law must hold. But it takes a great deal of careful argument, and relies on the reader having some experience of smooth, polished surfaces and near perfect spheres. Another non-Newtonian intuition which every child has is that a released slingshot travels radially outwards, rather than tangentially. Other examples can be found.

If an object is moving, there are only five possibilities. It may be *falling*; or it may be being *pulled or pushed* by something; or it may be moving *itself* along, expending effort as it does so (and therefore cannot be a passive object), or it may be *sliding* (by which I mean to include sliding on a slippery surface or down a slippery slope); or it may be *rolling*, in which case it must either be or have rollers or wheels. The last two cases can keep on going for a while after all effort has ceased. (We would call this phenomenon *coasting*: a gesture to Galileo.) This list does not include rotary or oscillatory motions: it is meant to cover all movements in the sense of *change of position*.

I believe there may be actually two distinct ways of conceptualising motion: as a displacement or as a trajectory. Displacement motion requires effort, and when the effort stops the motion stops: and it is characteristically under constant servo-control relative to position, i.e. it is conceptualised as *change of position*. Trajectory motion has inertia, keeps going unless stopped, (when there is an impact), is characterised as smooth motion *along a path*. It requires effort to start up, to stop or to alter, but less effort, or no effort, to maintain. Examples

are a thrown projectile, a car, sliding on ice, jumping. Displacement motion is Greek, trajectory motion is Galilean. Concepts such as aiming, impact, velocity (as a measure space), acceleration, are connected with the latter: concepts such as going, coming, dodging, avoiding, movement towards, away, are connected with the former. Both displacements and trajectories are histories, but the former are, essentially, merely transitions from their beginnings to their ends, which are positions of an object, whereas the latter have a definite *shape*: they can be extrapolated in time, for example, whence the concept of aiming. Falling, slidings, rollings and jumpings are examples of trajectories.

Forces can be transmitted in various ways. A rigid body can transmit a push, a string can transmit a pull. Luger and Bundy (1977) consider this mini-cluster in some detail.

7.10 Liquids

Liquid substances pose special problems, since, unlike pieces of solid stuff, 'pieces' of liquid are typically individuated *not* by being a particular piece of liquid, but by being in a particular *place* (a lake), or in some special relationship to a solid object (inside a cup). In Hayes 1978b I enlarge on an approach to solving these problems.

8. SOME STRUCTURED FORMALISATION TECHNIQUES

Constructing axiomatic formalisations which are 'heuristically adequate' (McCarthy & Hayes 1969) is an art, rather as writing good programs is an art. It is not yet very well developed (and indeed, one of the main aims of tackling naive physics is to gain skill in this relatively unexplored area), but a few stylistic points seem to be emerging.

One is the importance of *taxonomies*: finite exhaustive lists of the various types or categories of a kind of thing or of the possible states of a thing. We have seen these emerging already: ways of being supported, kinds of physical states, possible states of a fluid, (there are six: contained, flowing, spreading, wetting, falling or flying). In each case we have a group of axioms with the following form:

$$\phi(x) \equiv \phi_1(x) \vee \ldots \vee \phi_n(x)$$
$$\phi_1(x) \supset T_1$$
$$\vdots$$
$$\phi_n(x) \supset T_n$$

where the T_i are theories of what these particular cases are like. Such exhaustive lists can be very useful in making inferences, by a generalisation of the graphical consistency-checking computations widely used in vision (cf. Mackworth 1977). Intuitively, the iff ($\equiv$) means that if all but one of the disjuncts can be ruled out, then the remaining one must be the case. If the collection (T_i) of sub-theories is appropriately structured, then this can be a powerful technique for obtaining short proofs. For example, in the case of *support*, we can rapidly infer that if a passive thing is held up by a string (and that's all), and the string

breaks, then it must fall: for there is nothing underneath (case 1), there is no liquid for it to float on (case 3), it isn't attached to anything (case 2), and it can't fly: so it must be supported. So (by the basic *support* axiom), it must fall (i.e. this moment is the beginning of a falling history). A similar way of arguing can be used to show that water which flows to the edge of a table will fall (rather than, say, keep going horizontally, or pile up at the edge).

It may be significant that such taxonomies have the syntactic form of a definition (of ϕ in terms of the ϕ_i), but do not serve the role of definitions since the 'defined' token already occurs elsewhere in the axiomatisation.

A second stylistic point concerns existence and comprehension axioms. As we pointed out earlier, axioms which establish the existence of entities are vital in a formalisation which is to have nontrivial models. Examples we have met include the spaces (places) defined by physical boundaries (rooms, insides of cups), or by various (metric *and* non-metric) coordinate systems, and the histories which ensue when various states obtain, e.g. the falling which inevitably follows a state in which an object is unsupported. In these, and no doubt other, cases, there will be *comprehension axioms* which assert the existence of the required entity, and its relationship to the entities already established (the space *between* the walls: or *behind* the door: the falling *after* (and below) the moment (and place) where the object loses its support, and so on).

The point I want to make here, however, is that these are all *restricted* comprehension axioms. We cannot take arbitrary pieces of 3-space or 4-space-time and treat them as individuals: only ones which are related in a describable way to entities we already have confidence in, as it were. This selectiveness in ontological commitment is one of the characteristic differences, I believe, between common-sense reasoning and 'hard' scientific or philosophical reasoning. Common sense's ontology is prolix – entities of all sorts, concrete and abstract (objects, materials, colours, spaces, times, histories, events,. . .) are used with scant concern for philosophical tidiness, little appeal to an underlying ontological simplicity (compare subatomic physics, for example, or even Chemistry's periodic table): and yet it is also very controlled: contrast Goodman's (1966) nominalism, or axiomatic set theory, or the comprehension axiom scheme of the typed λ-calculus. The effect of both of these differences is to give a far more richly structured collection of entities: *fewer* than in these 'uniform' formalisations, but of far more *sorts*, and with a much richer collection of *kinds of relationship* between them.

One further point here: the use of global metric coordinate frameworks essentially restores the unrestricted comprehension by the back door. For, by using suitable coordinates, we can describe *any* piece of 3-space (air traffic corridors, for example, which have no physical boundaries at all), or *any* piece of space-time or *any* piece of fluid (e.g. the cubic centimetre whose top north-east corner was 5 cm below the surface at such-and-such a place in a certain river, at 19:30:06.8 hours on the 24 May, 1962). The resulting ontological *freedom* and uniformity is probably one reason why coordinate systems are so useful in (real) science.

9. WHY IT NEEDS TO BE DONE

I take it as obvious that the construction of a program which can be said to have common sense must involve, ultimately, the formalisation and common-sense knowledge such as naive physics in some way or other (and also, of course, naive psychology, naive epistemology, etc.). While there are those who would disagree with this – those, for example, who believe that a simple uniform learning procedure might eventually exhibit intelligence – their pre-theoretical assumptions differ so much from those made by most AI workers that it is better, I think, to regard such work as belonging to an essentially different field. I shall, in any case, not discuss this particular issue further here.

There is real room for disagreement, however, on methodology. The most dominant view within AI seems to be that it is necessary to construct a 'complete' program in order to demonstrate that one's ideas on representation are feasible. Working systems which exhibit an impressive total behaviour are taken to be the ultimate criterion of success. So strong is this requirement, indeed, that in many centres it is difficult for a student to obtain a PhD unless he has implemented an impressive working program. The naive physics proposal, as outlined here, deliberately avoids the construction of such a complete program. Our aim is to construct a formalisation which defines a *heuristically adequate* search space of possible inferences. Questions of how, exactly, to search this space; of controlling an interpreter; of information retrieval and relevancy – what might be called computational questions – as well as questions of the choice of data structures; how to implement fast searches; the choice of programming language – what might be called implementational questions – all will be deliberately ignored.

Relatively few workers in AI are adopting a similar methodological position, and yet I believe that it is vital to adopt it, in order to make substantial progress on representational issues. McCarthy (1977) makes some similar arguments.

It is not just a question of diverting resources, although this is an important issue. The more fundamental reason is, that quick success in constructing a complete working program seems necessarily to involve making simplifications and restrictions which make it impossible to tackle the essential representation problems. This happens in two ways.

Firstly, to achieve success in making an AI performance program, one must choose the domain of the program very carefully indeed. To be tractable, it must be restricted in some way, usually fairly drastically. A standard style of restriction is to limit the scope of the program: a restricted subject-matter for a reasoning program, a restricted vocabulary for a micro-world for a natural-language program, a restricted range of see-able objects for a vision system, etc. It follows, then, that the representation needed by the program is one which is not particularly *thorough*, in the sense used earlier. Moreover, the representation often can (and often does) rely on this restricted scope in using techniques which work for small 'toy' worlds, but which are simply inapplicable to more thorough uses. I have noted several examples above.

Secondly, computationally effective representations tend to be of rather low density. There is a good reason for this: a dense representation necessarily defines a large and explosively expanding search space of possible inferences. *If* the only heuristic devices available for controlling an inferential search are weak and general (numerical) heuristics, depth-first search as in MICROPLANNER local procedure invocation as in KRL-0, etc.), then effective computational behaviour cannot be achieved with such a search space. But these weak, general methods are essentially the only ones we know: hence, to be computationally effective, we must have sparse representations.

These two pressures, then, taken together, encourage the construction of sparse representations of limited scope which are tailored carefully to the particular desired behavioural repertoire of the task domain chosen for the performance of the program. But, as I have argued, thoroughness and density are essential properties of a representation which can be said to capture the meaning of common-sense knowledge adequately.

This is a methodological point, but there is a closely related point concerned with adequacy. A popular view in AI is that only the adequate performance of a program can be a criterion of success of an AI theory. (Indeed, I suggest elsewhere (Hayes 1978d) that this criterion is exactly what distinguishes AI from 'information processing psychology'). Accepting this, it is only a small step to accepting a sort of behaviourist criterion of adequacy for a representation: viz, that it support an adequate behaviour in some performance program. This criterion would accept a sparse, limited representation over a dense, thorough one, given the present state of the implementer's art. If the program *works*, so the argument would go, then its representation must adequately capture the intended meanings: for that is what we *mean* by 'adequate'.

The problem with this position, at least in its simpler versions, is that it takes no account of scale effects. One cannot make a program which behaves adequately in a large 'world' by any simple process of adding together programs which perform adequately in a number of smaller sub-worlds of that world. At least, this is so far the everyday world of common-sense. The world just doesn't split up that neatly: one needs the interaction between the parts as well. Thus, a representation of, say, the blocks world which is adequate for reasoning merely about blocks, on this criterion, will be less adequate for reasoning about blocks in the context of liquids, strings, rods, friction, pulleys, etc. And the pressures just noted, towards tailoring representations of limited scope, militate against what might be called upward compatibility of formalisations. Thus, even accepting the criterion as an ultimate test of an AI theory, which I do, I would still argue that applying it too rigorously too soon (typically, after 3 years work) is self-defeating. *We are never going to get an adequate formalisation of common-sense by making short forays into small areas, no matter how many of them we make.*

It might be objected that if all this is true, then a dense, thorough formalisation *cannot* be part of an effective AI program. But this is false. I have argued that weak, general methods of control are unable to handle dense, thorough

formalisations: clearly, we need more powerful control methods. There is one idea which I introduced in Hayes (1973) – see also Kowalski (1977), Pratt (1977), McDermott (1976), Davis (1976) – which suggests how to achieve the kind of power needed: to regard control not as a problem of defining a mechanism, but as itself a representational problem. We need to formalise the knowledge of *how* to make inferences, as well as the knowledge of the world which makes inferences possible. This meta-information can itself take part in the reasoning process, but also has a different and special relationship to the deductive interpreter: it describes its activity, rather than merely being grist for its mill. The development of formalisms, and associated interpreters, for expressing such meta-knowledge seems to me to be one of the most important tasks facing AI. But – and this is the crucial point for the present argument – the structures of both the formalisms and the meta-formalisations expressed in it, depend upon those of the world-knowledge formalism and the formalisations expressed in it. We cannot develop meta-formalisations in a vacuum (unless they be merely formalisations of the weak, general heuristics we already have); we must first have some realistically complex examples of common-sense formalisations, whose deductive properties will be described in the meta-formalisation.

10. WHY IT CAN BE DONE

A different 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 to make the attempt and succeed, so all I can do here is to convey my reasons for feeling optimistic. There are four.

The first is based on my recent experiences in tackling the 'liquids' problem, which I have long believed was one of the most difficult problems in 'representation theory' (Hayes 1975). The idea of quantifying over pieces of space (defined by physical boundaries) rather than pieces of liquid, enables the major problems to be solved quite quickly, to my surprise. The key point here was finding the correct procedure for *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. McCarthy (private communication 1977) has, for example, begun a new approach to epistemic formalisations based on the individuation of 'concepts', i.e. thoughts in people's heads.

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 will 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, heuristically 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 there is an obvious methodology for getting it done,

and this methodology has, in recent years, proved very successful in a number of areas.

11. HOW TO GET IT DONE

There is a tried and true way of getting knowledge out of people's heads and into a formalisation. Within AI, it has been called 'knowledge engineering' by Feigenbaum (1977); 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: indeed, the expert is usually the linguist himself. 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. (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.) It seems to be sound to have several 'experts' involved, as it is easy to miss some obvious distinctions when working alone.

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 so 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, the formalisations need be little more than carefully-worded English sentences. One can make considerable progress on ontological issues, for example, without actually *formalising* anything. Fairly soon, however, it will be necessary to express the intuitions formally. Here, I think one should be liberal in allowing a free choice of formal language. Many people find frame-like notations agreeable: others like semantic networks, etc. There is no reason why such superficial variants of first-order logic, or even more exotic formalisms, should be banned: 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. A more serious point is that particular clusters may suggest special ad-hoc representations. Shapes may be represented diagrammatically, for example. One can imagine a cluster, represented in some idiosyncratic way, whose internal inferential relationships were inaccessible from outside, but which was interfaced to the rest of the formalisation by a defined translation of part of itself into the reference formalisation (first-order logic): say statements of relative position and orientation. It will be difficult to prevent such things happening, and maybe one should not try. But there are grave dangers, since this way of proceeding prejudges the possible interactions in the formalisation as a whole, and this may encapsulate a serious mistake in a way which will be hard to detect and even harder to rectify.

There are several other ways to find concept-clusters. For example: looking in a thesaurus; choosing a particular domain (cookery, volumetric measurements of various substances) and attempting to describe it; analysing some everyday act in detail (e.g. spreading a sheet over a bed by holding two corners and flicking: why does that work?). I expect these, and others, will be useful starting-points.

12. IS THIS SCIENCE?

It will be objected that to attempt to formalise knowledge in the abstract, i.e. divorced from particular sensory modality or task domain, is unscientific because there are no clear criteria of success or failure. What would it be like to fail? If this question cannot be answered, then naive physics is mere literary criticism.

I think this objection is well taken and needs a more adequate reply than I am currently able to give. The point is, that one can always get *somewhere* with a formalisation: who is to say that where one has got to is not far enough? One can only say, I think, that people's common intuition is the guide. If there are some 'obvious' physical facts which cannot be made to follow 'easily' or 'naturally' from the axioms, then more work has to be done. One can apply the usual scientific judgements of 'elegance', 'economy', etc. to compare rival formalisations. (All the quoted words cry out for further discussion, which I will not attempt here.) It is worth remarking that linguistics is in *exactly* the same position, and regularly squirms on the methodological hook: judgements

of physical plausibility and elementary causality are certainly as reliable as judgements of grammaticality by native speakers; and indeed are largely independent of cultural and linguistic boundaries, so are probably rather more reliable as source data. (Which may suggest that we should borrow the competence/performance distinction to protect ourselves from behavioural refutation: but there are deep problems indeed about trying *that* trick.)

It would be nice if naive physics could be linked more closely with the mass of data now available in Piagetian psychology on the development of physical concepts during childhood (although none of this is uncontroversial, it seems). Certainly, *compatibility* with this data should be a constraint on naive-physics formalisation. It is however a very weak constraint, since the data is compatible with many different developmental theories, and is not usually sufficiently detailed to distinguish one from another (see Prazdny 1978 for some comments in this direction). I would hope that the construction of a naive physics might show up some new mechanism of developmental change in conceptual frameworks. To some extent this is happening already, since building a formalisation is often a matter of *developing* (in exactly the right sense) the partial formalisations one already has.

ACKNOWLEDGEMENTS

This paper was written during a sabbatical visit to the Institut pour les etudes sémantiques et cognitives, Geneva. I am grateful to the Directrice, Mme M. King, for inviting me there; and to all members of the Thursday seminar, especially Maghi King, Giussepe Trautteur and Henri Wermus. Conversations with Mimi Sinclair on Piagetian research were also of great help. My wife, Jackie, typed several drafts of the manuscript and was an unfailing source of reliable, sound, common-sense intuitions.

REFERENCES

ANDERSON et al., 1972. Beyond Leibnitz. *Memo AIM*, Stanford Artificial Intelligence Project.

BINFORD et al., 1976. Computer Integrated Assembly Systems. *MEMO AIM-285*, Stanford Artificial Intelligence Project.

BOLLES, 1976. Verification Vision within a Programmable Assembly System. Stanford AI Memo No. 295, 1976.

BUNDY, 1978. Exploiting the Properties of Functions to Control Search. (To appear.)

DAVIS, 1976. Applications of Meta-level Knowledge to the Construction, Maintenance and Use of large Knowledge Bases. *HPP MEMO 76-7*, Stanford University.

FEIGENBAUM, 1977. Themes and Case Studies of Knowledge Engineering. *Proc. 5th IJCAI Conference*, M.I.T.

GOODMAN, 1966. The Structure of Appearance. New York, Bobs-Merrill Co.

HAYES, 1971. A Logic of Actions. *Machine Intelligence 6*, Edinburgh University Press.

HAYES, 1973. Computation and Deducation. *Proc. 2nd MFCS Sympsoium, Czechoslovakian* Academy of Sciences.

HAYES, 1975. Problems and Non-problems in Representation Theory. *Proc. 1st AISB Conference*, Sussex University.

HAYES, 1977. In Defense of Logic. *Proc. 5th IJCAI Conference*, M.I.T.

HAYES, 1978a. The Logic of Frames. *The Frame Reader*. (To be published by De Gruyter, Berlin.)

HAYES, 1978b. Naive Physics: Ontology of Liquids. *Working Paper 35*, Institute for Semantic and Cognitive Studies, Geneva.
HAYES, 1978c. Naive Physics II: Histories. (In preparation.)
HAYES, 1978d. On the Difference between Psychology and Artificial Intelligence. *AISB Bulletin*. To appear.
KOWALSKI, 1977. Algorithm = Logic + Control. Memorandum, Imperial College, London.
LANDIN, 1970. A Program-Machine Symmetric Automata Theory. *Machine Intelligence 5*, Edinburgh University press.
LUGER & BUNDY, 1977. Representing Semantic Information in Pulley Problems, *Proc. 5th IJCAI Conference*, M.I.T.
MACKWORTH, 1977. Consistency in Networks of Relations. *Artificial Intelligence, 8.*
McCARTHY & HAYES, 1969. Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Machine Intelligence 4*, Edinburgh University Press.
McCARTHY, 1977. Epistemological Problems of Artificial Intelligence. *Proc. 5th IJCAI Conference*, M.I.T.
McDERMOTT, 1976. Flexibility and Efficiency in a Computer Program for Designing Circuits. *PhD thesis*, MIT Artificial Intelligence Lab.
McDERMOTT, 1977. Artificial Intelligence and Natural Stupidity. *SIGART Newsletter.*
PRATT, 1977. The Competence-Performance Distinction in Programming. *Proc. 4th ACM Symposium on Principles of Programming Languages*, Los Angeles.
PRAZDNY, 1978. Stage two ofthe Object Concept Development: a Computational Study. *Memorandum*, Essex University.
SCHANK, 1975. Conceptual Information Processing. North-Holland.
ULLMAN, 1977. The Interpretation of Visual Motion. *PhD thesis*, M.I.T.
WILKS, 1975. A Preferential, Pattern-Matching Semantics for Natural Language Understanding. *Artificial Intelligence, 6.*
WILKS, 1977. Good and Bad Arguments about Semantic Primitives. *Memo 42*, Artificial Intelligence Dept., Edinburgh University.
WINOGRAD, 1972. Understanding Natural Language. Edinburgh University Press.
ZEEMAN, 1962. The Topology of the Brain and Visual Perception. *Topology of 3-Manifolds* (ed. K. Fort), Prentice-Hall.