25

A Logic of Actions

P. Hayes
Metamathematics Unit
University of Edinburgh

THE FRAME PROBLEM

One of the central principles upon which intelligent devices seem to operate is that of maintaining internal models of their external environments. In artificial systems which have been constructed to date various representations for this internal model have been used; but in every nontrivial case the need arises to consider the effect, upon the structure of the model, of the performance by the system of *actions* in the external world, so that their potential consequences may be reckoned.

How difficult this is, depends upon both the complexity of the model and its method of representation. In particular, it is usually easy when the problem is posed in the classical heuristic search paradigm, and the data structures used to represent static configurations of the puzzle are relatively unproblematic (arrays, lists, and so on). For in this case [*see*, for instance, Manna (1970) and Fikes (1970) for examples] one can use the ordinary device of *assignment* to model the changes in the world which result from the performance of actions. The lack of side-effects reflects the simplicity of the physics which such models embody. This limitation to elementary forms of interaction is not, of course, intrinsic to the heuristic search method; but when more complex models are constructed it becomes less trivial to pursue the consequences of performing an action. The use of assignment to portray the doing of actions *does* seem to presuppose a trivial physics.

Another method of constructing microcosms is to use a logical language to *describe* the real world (McCarthy 1959, McCarthy and Hayes 1969). This approach is more general than the heuristic search method (but the latter – when it has sufficient expressive power – wins at present by its computational advantage). The key idea is to use expressions denoting *situations* to separate out assertions according to which (static) state of the world they purport to describe. Assertions mentioning several different situations can then be used to describe dynamical laws which move us from one situation to another. This

use of situations ensures that we do not fall into confusing an assertion made on one occasion with a similar assertion made about a different state of affairs; and this is clearly desirable. But in some ways the resulting sharp separations between states of affairs are an embarrassment. For if we distinguish two situations $\mathbf{s}_1$ and $\mathbf{s}_2$, then from the fact, if such it be, that a predicate **p** is true of $\mathbf{s}_1$, nothing whatever follows concerning $\mathbf{s}_2$. And this is true even when $\mathbf{s}_2$ is directly associated with $\mathbf{s}_1$. Say $\mathbf{s}_2$ results from $\mathbf{s}_1$ by the performance of some action: $\mathbf{s}_2 = do(\mathbf{a}, \mathbf{s}_1)$ then no matter how remote – speaking intuitively – the connection between the property **p** and the action **a**, it still does not follow that **p** is true of $\mathbf{s}_2$. If we want it to so follow we must state this explicitly. Now, unfortunately, there are innumerable facts which might remain unchanged when actions are performed. So instead of writing a 'law of motion' in the form $\mathbf{A}(\mathbf{s}) \supset \mathbf{B}(do(\mathbf{a}, \mathbf{s}))$ where **A** and **B** are fairly short expressions, we are apparently obliged to list systematically all conceivable facts which are *not* changed. So that the law looks more like

$$(\mathbf{C}_1(\mathbf{s}) \& \ldots \& \mathbf{C}_n(\mathbf{s}) \& \mathbf{A}(\mathbf{s})) \supset .$$
$$\mathbf{C}_1(do(\mathbf{a}, \mathbf{s})) \& \ldots \& \mathbf{C}_n(do(\mathbf{a}, \mathbf{s})) \& \mathbf{B}(do(\mathbf{a}, \mathbf{s}))$$

for some very large *n*. This works for small problems (such as the familiar hungry anthropoid), but these are usually better formalized in the heuristic search paradigm anyway. It is clearly going to become impractical in any elaborate system.

That is the frame problem. Several points can be made. *First*, it should be clear that the underlying problem is not peculiar to the 'logical language' approach to model-building. Rather, it is a fundamental difficulty – delineating precisely the ways in which actions affect a complicated world – which appears in the above guise in linguistic models, but which appears everywhere in some form.

Secondly, it is not, contrary to opinion, just an implementation difficulty. The problem is not *simply* that long laws give rise to inefficiencies in the theorem-proving process (although of course they do, and I would not want to disparage such an argument against them). The problem runs much deeper. First of all, it is not clear whether we could in all cases write down adequate 'long laws'. Whether or no some assertion is affected by some action may not be determinable once for all. It may depend upon a detailed analysis of the situation; that is, in the present context, upon a long deduction. It is not clear how such deductions could proceed within a first-order (even modal) logic. Moreover, this method of getting round the difficulty is essentially *static*. Suppose we had a set of 'long laws' which were adequate for the theory so far. Now suppose a new predicate (say) is introduced. Immediately all the old laws become inadequate: for they do not specify what happens to assertions involving the new predicate when actions are performed.

Thirdly, the frame problem is distinct from what may be termed the *prediction problem*; that is, that any prediction which the system makes is liable to be subsequently contradicted by its immediate experience. Thus

some rational way of controlling the editing and updating of its belief structure is needed, and this can become very difficult to manage properly. This problem has been confused with the frame problem, perhaps because it arises from inadequacies in the world-model, and the frame problem often gives rise to such inadequacies. But they do also have a curious relationship, as follows. Suppose we decided to end the frame problem for ever by decreeing that *all* properties are unchanged by *all* actions, except those explicitly mentioned in the laws of motion (that is, the 'short' laws). Then it is extremely likely that the resulting logic would rapidly generate inconsistencies, for changes that were forced to occur would yield consequences which contradicted the blanket assumption. Now if one regards this assumption as a *prediction* on the meta-level (a prediction that it is consistent to assume that such changes will not take place), then finding a direct contradiction seems analogous to making a direct observation which condemns some prediction about the world. Viewed in this way, the blanket assumption of no change is a (rather simplistic) model *of the system itself.* The system plays the role, for *this* model, of the real world. While this observation is rather interesting, I shall not explore it further, other than to remark that a good solution to the prediction problem would clearly yield a partial solution to the frame problem [this is the sort of approach mentioned in Part 4 of McCarthy and Hayes (1969)].

Several partial solutions to the frame problem have been suggested. One is the use of *frames* (McCarthy and Hayes 1969). A frame is a classification of statements into groups which are independent in the sense that an action may alter members of one group without affecting any of the other groups. For instance, statements about colours can be put in a separate group from statements about location. Unfortunately the classification can in general only be rather coarse – as the above example indicates. Thus this goes only a little way towards solving the problem.

There are simple situations where a frame can be used to good effect. These are precisely the models which can readily be put into the form described earlier, so that the effects of actions are easily described by assignment statements. This brings us to the second partial solution to the frame problem: to consider only sufficiently simple world-models that the assignment method might always be adequate. This is indeed very popular just now, since many of the classical puzzles can be described this way. It underlies all the projects which propose the heuristic search paradigm, or its latest variant, the non-deterministic algorithm (Foster and Elcock 1969, Fikes 1970, Manna 1970) as the model representation. These all assume an array-like storage system and are thus wedded to the assignment method of representing change. But, as I have argued above, this is not going to be adequate for more complex domains. Consider, for instance, a cup on a saucer. If we move the saucer, the cup moves too: but if we move the cup, the saucer stays where it is. It is not difficult to invent arbitrarily complicated examples of this kind.

There are other suggested solutions or partial solutions in the literature, but none of them seem to be capable of direct use at present (Minsky 1961, McCarthy and Hayes 1969, Part 3).

The weakness of all these methods is that they are too inflexible. They do not allow interaction between the details of the world-model and the changes resulting from the performance of actions. Thus they impose rigid classifications which are therefore either weak, or else admit only a trivial physics. It is the aim of this paper to provide a more flexible interface between the physics of the world-model and the formal behaviour of the logic.

ACTIONS AND CAUSALITY

The basic idea is this. The only way of moving from one situation to another is by performing an action. To each action there corresponds a certain (small) set of individuals, those which are *directly affected* by the action. When the action is performed, these individuals associated with it are liable to change their properties. It would be pleasant if we could assert that all other individuals have all their properties unchanged: but this is to assume too much. For individuals may be connected together in all sorts of ways, and actions may touch off long chains of cause and effect. *We shall assume that this causal connection between individuals is axiomatized* by a binary relation written '$\rightarrow$'. Thus $(\mathbf{a}\rightarrow\mathbf{b})$ is to mean that some property of (the value of some predicate or function applied to) **a** is causally related to some property of **b**, so that if the latter changes the former is liable to.

The utility of this is that when we can prove that $\sim(\mathbf{a}\rightarrow\mathbf{b})$, then we know that *no* change to **b** will cause any change to **a**. This immediately suggests a rule of inference which can take advantage of such facts. For if we know that **a** does not bear the $\rightarrow$ relation to any of the individuals directly affected by an action, then we can infer that all properties of **a** are preserved during the performance of the action.

This approach to the frame problem makes few presuppositions. The two most important are that the world is *deterministic*, that is, changes do not occur spontaneously; and that there is only one agent in the universe.

This latter constraint is not essential and the theory could be elaborated to allow several agents: but with the important proviso that we have at present no idea how to handle the *simultaneous* performance of actions by two or more individuals.

Before progressing to details it seems appropriate to mention some further weaknesses. I do not consider any of the problems which arise when one allows actions to have the structure of *programs*, with loops, and so on (McCarthy and Hayes 1969). Also I do not consider the problems arising from attempting to mix talk of actions with tensed statements; nor with any other modal operators. The whole theory will be couched in ordinary first-order logic. This is deliberate, as these extras would make the whole system very complicated and would obscure the point. And in any case, I do not

know how best to handle these problems. In particular, the whole field of computation theory seems to be very active at present, and I imagine that most of the current theories could be adapted to the present system.

These matters will be taken up again later.

LOGICAL CONVENTIONS AND NOTATION

In general we shall follow the version of first-order logic described in Schoenfield's textbook (1967), except that our theories will be sorted (*see* below). We shall also follow Schoenfield's notation, so that **A, B, C,** etc. denote arbitrary statements, **u, v,** arbitrary wffs; **p, q,** predicate symbols; **f, g,** function symbols; **e,** constant symbols (that is, function symbols of degree 0); **x, y, z,** variables; **a, b, c, d,** terms. Also **h** denotes an arbitrary nonlogical symbol. We shall, however, use the symbols $\sim$, $\vee$, &, $\supset$, $\equiv$ for the logical connectives, reserving the symbol $\rightarrow$ for the causality relation.

We shall also use the Church dot convention and the usual theorem-proving terminology: an *atom* is a predicate with arguments; a *literal* is an atom or its negation; a *ground* wff is one without variables. An occurrence of a term (with range $\{s\}$: *see* below) in an expression is *maximal* if the term does not occur as a proper subterm of any term (with the same range) occurring in the expression.

The expression $\mathbf{u}_{\mathbf{x}}[\mathbf{a}]$ denotes the result of substituting **a** for **x** in **u**. We may omit the **x** when no confusion will result, so that **u**[**a**] denotes an expression containing **a** in which all occurrences of **a** are distinguished.

A *theory* T is (as usual)

(1) a finite set of nonlogical symbols (predicate and function symbols).

(2) a denumerable set of nonlogical axioms, each of which is a wf statement written using the nonlogical symbols and the equality symbol $=$.

(3) the set of logical axioms appropriate to the vocabulary.

A theory T is *sorted* when there is

(1) a finite set S_T of sorts.

(2) for each nonlogical symbol **h** of T, a partial function $\mathbf{h}^0$ of the same degree from S_T to S_T.

(3) a total function $=^0$ from $S_T \times S_T$ to S_T.

Suppose ϕ is a mapping from variables to sorts: then clearly we can extend ϕ inductively to a partial mapping ϕ^0 from terms and atoms to sorts. As ϕ varies through such mappings, $\phi^0(\mathbf{a})$ varies through S_T. Let **a** be a term occurring in **u**, then the *range of* **a** *in* **u** is the subset of S:

$\{\phi^0(\mathbf{a}) \mid \phi$ is a mapping from variables to sorts and $\phi^0(\mathbf{v})$ is defined for every term or atom **v** occurring in **u**$\}$.

The *range* of a term **a** is the range of **a** in **a**. Clearly, the range of a ground term is a singleton.

In the usual recursive definition of a wff we insist that every subterm has a nonempty range in the wff. If $Q\mathbf{x}\,.\,\mathbf{u}$ is a wff, where Q is some quantifier,

then the quantification is understood to be relativized to the range of $\mathbf{x}$ in $\mathbf{u}$.

The idea is that each function symbol $\mathbf{f}$ has certain combinations of sorts allowable for its arguments, and the sort of its value is a function of the sorts of its arguments. When a combination is not allowable we make $\mathbf{f}^0$ be undefined: hence the $\mathbf{f}^0$ are partial functions. In the case of predicates, we do not need to specify the sort of the value, so we say that a set of arguments are of sorts acceptable to $\mathbf{p}$ just in case $\mathbf{p}^0$ is defined, regardless of its actual value.

This notion of sorts is the most general one I have seen. Usually we shall not need to use its full power.

There is no difficulty in mechanizing such a sort structure provided that the functions $\mathbf{h}^0$ are somehow conveyed to the machine in an easily computable form. Also, it does not materially affect the formal behaviour of the logic. The use of the expression $\mathbf{u}_{\mathbf{x}}[\mathbf{a}]$ will be taken to imply that the wff being mentioned is well formed, as this is no longer automatic. The range of a term $\mathbf{b}$ in $\mathbf{u}_{\mathbf{x}}[\mathbf{a}]$ is a subset of its range in $\mathbf{u}$. We need to add axioms enabling us to infer that $\sim(\mathbf{a}=\mathbf{b})$ whenever the ranges of $\mathbf{a}$ and $\mathbf{b}$ in the equality are disjoint. This can be done by using classifying predicates, that is, for each $s \in S$, a unary predicate symbol P_s, such that P_s^0 is true everywhere. P_s is to be true just when its argument has sort s, but of course this is a constraint upon the semantics. We shall assume that our theories contain classifying predicates.

The P_s are logical symbols. However it is clear that to any sorted theory with classifying predicates there corresponds an unsorted theory in which the classifying predicates are ordinary nonlogical symbols. The unsorted theory will in general be much *larger* (more symbols and more axioms) than the sorted theory.

Semantics

A *structure* for T is:

(1) a nonempty set $|\mathscr{A}_s|$ for each $s \in S_T$. We define $|\mathscr{A}| = \cup |\mathscr{A}_s|$ as the *universe* of $\mathscr{A}$.

(2) for each function symbol $\mathbf{f}$ of T, with degree n, a partial function $\mathbf{f}_{\mathscr{A}}$ from $|\mathscr{A}|^n$ to $|\mathscr{A}|$.

If $x_i \in |\mathscr{A}_{s_i}|$ for $1 \leqslant i \leqslant n$, then $\mathbf{f}_{\mathscr{A}}(x_1, \ldots, x_n) \in |\mathscr{A}_{\mathbf{f}^0(s_1, \ldots, s_n)}|$ provided $\mathbf{f}^0(s_1, \ldots, s_n)$ is defined; otherwise $\mathbf{f}_{\mathscr{A}}(x_1, \ldots, x_n)$ is to be undefined.

(3) for the equality symbol, the predicate $=$ which is true when its arguments are equal and false otherwise.

(4) for each classifying predicate P_s, the monadic predicate P_s which is true in $|\mathscr{A}_s|$ and false everywhere else in $|\mathscr{A}|$.

(5) for each predicate symbol $\mathbf{p}$ of T, with degree n, a partial predicate $\mathbf{p}_{\mathscr{A}}$ on $|\mathscr{A}|^n$. If $x_i \in |\mathscr{A}_{s_i}|$ for $1 \leqslant i \leqslant n$, then $\mathbf{p}_{\mathscr{A}}(x_1, \ldots, x_n)$ is defined iff $\mathbf{p}^0(s_1, \ldots, s_n)$ is.

We shall adjoin to T, following Schoenfield, a new constant for each element of $|\mathscr{A}|$, called the *name* of the individual (**i** and **j** will denote names). The range of **i** is to be $\{s\}$ when the individual named by **i** is in $|\mathscr{A}_s|$.

With these modifications, the rest of the theory is as in Schoenfield (1967, p. 19). The above definition of the range of a name ensures that the quantifiers receive the proper meaning. In the recursive definition of the truthvalue of a closed formula one must remember that $\mathbf{B}_x[\mathbf{i}]$ must be well formed.

The usual idea of the value $\mathbf{u}_{\mathscr{A}}$ of a closed wff **u** is assumed. It is easy to show that $\mathbf{u}_{\mathscr{A}}$ is always defined and belongs to $|\mathscr{A}_s|$, where $\{s\}$ is the range of **u**. We also assume the notions of $\mathscr{A}$ being *valid in* $\mathscr{A}$, $\mathscr{A}$ being a *model* of T, (every axiom of T is valid in $\mathscr{A}$), and **A** being *valid* (**A** is valid in every $\mathscr{A}$).

A detailed account of first-order sorted logic is in preparation.

STATIC AND KINEMATIC THEORIES

A sorted first-order theory T will be called a *static* theory just in case it has a binary predicate $\rightarrow$ and the following axioms

$$\forall x(x\rightarrow x) \qquad (\rightarrow 1)$$

$$\forall xyz\,.\,((x\rightarrow y)\&(y\rightarrow z))\supset(x\rightarrow z) \qquad (\rightarrow 2)$$

That is, $\rightarrow$ is a quasi-ordering.

A static theory is intended to provide the means of describing a fixed state of affairs. There are not to be any situations mentioned in the statics language. Situations and actions will be introduced into the *kinematic* theories, described below.

No constraints are imposed upon the statics theory other than it provide statements of causality. This causality relation will be used by the kinematic theory to infer that certain actions do not change the properties of certain objects.

Bearing in mind that $(\mathbf{a}\rightarrow\mathbf{b})$ means that some property of **a** is liable to change iff some property of **b** changes, it is I think obvious that axioms $(\rightarrow 1)$ and $(\rightarrow 2)$ are intuitively true. One might ask whether any further conditions can be imposed upon the relation. In figure 1 we illustrate counter-examples to two of the more plausible of such conjectures. These are both constructed in a world of toy building blocks.

Let T be a static theory, and let Ψ be the set of nonlogical symbols of T. Let Θ be a nonempty subset of Ψ. We shall define the notion of a *kinematic extension* $K\,(=K_\Theta(T))$ *of* T *on* Θ (or simply a *kinematic theory*).

First, the sorts of K are to be those of T plus the two new ones, *situation* and *action*. The logical symbols of K are to be those of T plus the classifying predicates $P_{situation}$ and P_{action}: we assume that equality is extended to the new sorts. Now, let **h** be a nonlogical symbol of T with degree n, but not $\rightarrow$. Then $\bar{\mathbf{h}}$ is a new nonlogical symbol of degree $n+1$ whose sort function $\bar{\mathbf{h}}^0$ is defined:

$$\bar{\mathbf{h}}^0(s_1,\ldots,s_{n+1})=\mathbf{h}^0(s_1,\ldots s_n) \text{ if } s_{n+1} \text{ is } \textit{situation};$$
$$=\text{undefined otherwise.}$$

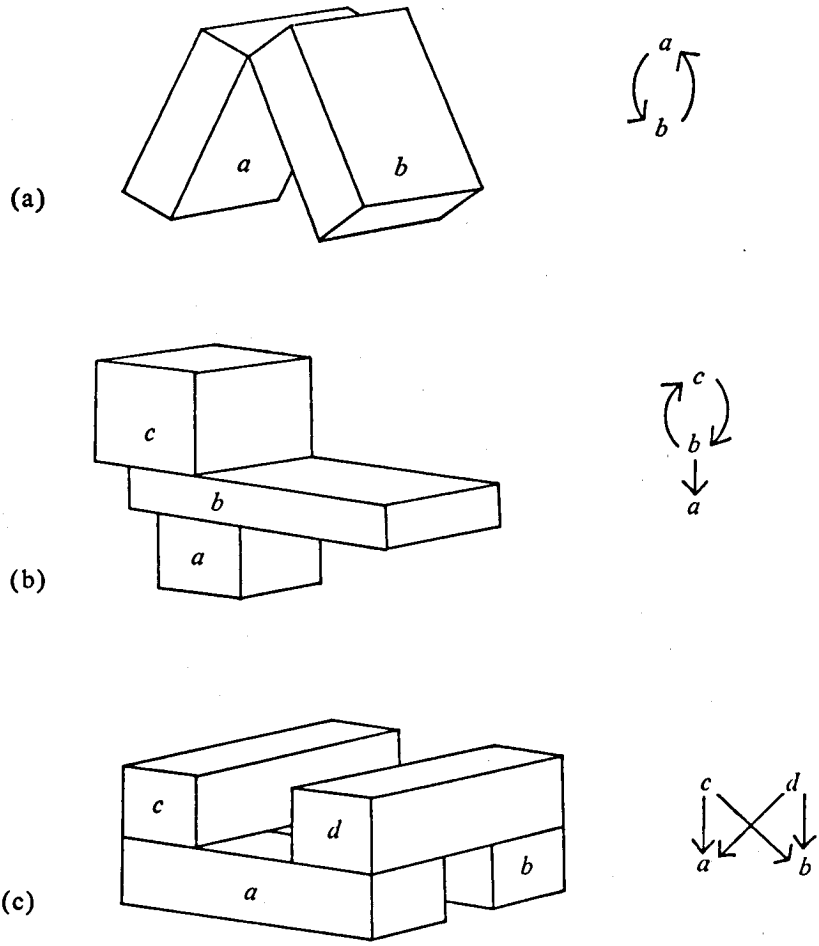

Figure 1. (a) $\rightarrow$is not a partial ordering $(((x \rightarrow y) \& (y \rightarrow z)) \supset x \rightarrow z)$; (b) another example; (c) $\rightarrow$ does not have sups or infs.

If $\mathbf{h}$ is $\rightarrow$, let $\tilde{\mathbf{h}}$ be a three-place predicate letter whose sort function is defined just when the first argument is an appropriate sort of T, the second argument an appropriate sort of T or *action*, and the third argument is *situation*. We shall use $\rightarrow$ for this new symbol also and write $\mathbf{(a \rightarrow b, c)}$ rather than $\rightarrow\mathbf{(a, b, c)}$.

Let Θ^s be $\{\tilde{\mathbf{h}} : \mathbf{h} \in \Theta\}$. Then the nonlogical symbols of K are

$$(\Psi - \Theta) \cup \Theta^s \cup \Phi$$

where Φ is some set of new symbols containing the binary function symbol *do* with

$$do^0(action, situation) = situation$$

$$do^0(s_1, s_2) = \text{undefined otherwise.}$$

If $\mathbf{h} \in \Phi$ we insist that $\mathbf{h}^0$ is undefined whenever any but its *last* argument is *situation*.

If **a** is a term of K whose range contains some sort other than *situation* or *action* we shall call **a** a T-term.

A term **a** occurs *crucially* in a wff **u** when **u** contains $\mathbf{h}(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{b})$, where **b** has range {*situation*} in **u**, and **a** is $\mathbf{h}(\mathbf{a}_1, \ldots, \mathbf{a}_n, \mathbf{b})$ or is some $\mathbf{a}_i$.

Suppose **u** is a wff of T containing some symbols in Θ, and let **a** be a term of K whose range includes *situation*. Then $\mathbf{u}[\![\mathbf{a}]\!]$ denotes the wff of K obtained by replacing every $\mathbf{h} \in \Theta$ occurring in **u** by $\bar{\mathbf{h}}$, and writing **a** in the final argument place of each such $\bar{\mathbf{h}}$. More generally $\mathbf{u}[\![\mathbf{a}]\!]$ will denote some wff of K in which every occurrence of a term with range {*situation*} in **u** is an occurrence of **a**.

The nonlogical axioms of K are to be

(1) Every axiom of T which is a wff of K.

(2) For every other axiom A of T, the axiom $\forall \mathbf{x} \mathbf{A}[\![\mathbf{x}]\!]$.

(3) A set of wff containing some symbols from Φ.

(4) The axiom *schema* ($\rightarrow$3):

$$\mathbf{A}[\![\mathbf{y}]\!] \supset .(\sim(\mathbf{a}_1[\![\mathbf{y}]\!] \rightarrow \mathbf{x}, \mathbf{y}) \& \ldots \& \sim(\mathbf{a}_n[\![\mathbf{y}]\!] \rightarrow \mathbf{x}, \mathbf{y})) \supset \mathbf{A}[\![do(\mathbf{x}, \mathbf{y})]\!]$$

where A is a literal and $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are all the T-terms which occur crucially in A.

This completes the definition of $K_{\bar{\Theta}}(T)$.

The set Θ comprises the names of all the properties whose value can be affected by the performance of actions. The set Φ may contain, for instance, predicates of actions and functions for constructing actions from objects or other actions. The inclusion of the axioms mentioned in (1) and (2) above ensures that if A is a theorem of T then either A or $\forall \mathbf{x} \mathbf{A}[\![\mathbf{x}]\!]$ is a theorem of K. This is the sense in which K is an extension of T. The axioms mentioned in (3) are intended to include whatever theory of actions is thought appropriate and also *laws of motion*, that is, assertions of the form:

$$(\mathbf{A}[\![\mathbf{a}]\!] \& \mathbf{B}[\mathbf{b}]) \supset \mathbf{C}[\![do(\mathbf{a}, \mathbf{b})]\!].$$

It will be by compounding such laws that plans are formed. As was remarked earlier, the only plans which can be constructed in this logic are simple unbranching sequences of actions. The schema ($\rightarrow$3) is the heart of the system. It allows us to infer that facts do *not* change when actions are performed.

Extending→3

It might be thought that the schema ($\rightarrow$3) was unduly restrictive in its insistence upon A being a literal. We shall rectify this at once.

It would be pleasant if we could allow A to be any statement, but this is not possible. The difficulty arises over negation. Let us say that A is in $\sim$ —*miniscope form* when it contains negation signs only in literals, and otherwise contains only the connectives & and $\vee$. Every A is equivalent to some

B in $\sim$ – miniscope form. It will turn out that ($\rightarrow$3) can be generalized to statements in $\sim$ – miniscope form, but that these statements undergo a certain transformation, as follows.

Let **a** and **b** be terms whose ranges contain respectively *action* and *situation*, and let **A** be a wff of K in $\sim$ – miniscope form. Then $\mathbf{A}^*_{(\mathbf{a},\mathbf{b})}$ is the wff which results when every quantifier $Q\mathbf{x}\ldots$ in **A**, where **x** is a T-term in **A**, is replaced by

$$Q\mathbf{x}((\mathbf{c}_1[\![\mathbf{b}]\!]\rightarrow\mathbf{a},\mathbf{b})\vee\ldots\vee(\mathbf{c}_n[\![\mathbf{b}]\!]\rightarrow\mathbf{a},\mathbf{b}))\vee\ldots$$

where $\mathbf{c}_1,\ldots,\mathbf{c}_n$ are all those T-terms which:

(1) occur crucially in the scope of Q;

(2) are free in the scope of Q; and

(3) contain **x**.

This is the transformation we wanted to define. Although the description is somewhat complicated the idea is simple: every quantification over objects (as opposed to situations or actions) is restricted to those objects which are not connected to the action **a** in situation **b**. The second and third conditions on the $\mathbf{c}_i$ ensure that scope conventions are not violated. If **A** is in $\sim$ – miniscope form, so is $\mathbf{A}^*_{(\mathbf{a},\mathbf{b})}$.

We can now state and prove the first consequence of ($\rightarrow$3).

Lemma 1

Suppose $\mathbf{A}[\![\mathbf{x}]\!]$ is in $\sim$ – miniscope form, and let $\mathbf{c}_1,\ldots,\mathbf{c}_n$ be all the T-terms which occur crucially and are free in **A**. Then

$$\mathbf{A}[\![\mathbf{x}]\!]\supset.\ \forall\mathbf{y}.((\mathbf{c}_1\rightarrow\mathbf{y},\mathbf{x})\vee\ldots\vee(\mathbf{c}_n\rightarrow\mathbf{y},\mathbf{x}))\vee(\mathbf{A}[\![do(\mathbf{y},\mathbf{x})]\!])^*_{(\mathbf{y},\mathbf{x})}$$

is a theorem of K.

The proof is by induction on the length of **A**. If **A** is a literal then the result follows directly from ($\rightarrow$3). If **A** is $(\mathbf{B}\,\&\,\mathbf{C})$ or $(\mathbf{B}\vee\mathbf{C})$ then the result follows by propositional calculus. So suppose A is $Q\mathbf{x}.\mathbf{B}$. Let $\mathbf{c}_1,\ldots,\mathbf{c}_n$ be all the T-terms which occur crucially and are free in **B**, and let $\mathbf{C}^j_i$ be $((\mathbf{c}_i\rightarrow\mathbf{y},\mathbf{x})\vee\ldots\vee(\mathbf{c}_j\rightarrow\mathbf{y},\mathbf{x}))$. Then the induction hypothesis is

$$K\vdash\mathbf{B}[\![\mathbf{x}]\!]\supset.\ \forall\mathbf{y}.\ \mathbf{C}^n_1\vee(\mathbf{B}[\![do(\mathbf{y},\mathbf{x})]\!])^*_{(\mathbf{y},\mathbf{x})}$$

where without losing generality we may take **x** and **y** distinct from **c**. Then by the distribution rule (Schoenfield, page 32)

$$K\vdash(Q\mathbf{z}\mathbf{B}[\![\mathbf{x}]\!])\supset.\ Q\mathbf{z}\forall\mathbf{y}.\ \mathbf{C}^n_1\vee(\mathbf{B}[\![do(\mathbf{y},\mathbf{x})]\!])^*_{(\mathbf{y},\mathbf{x})}$$

Now suppose without loss of generality that $\mathbf{c}_1,\ldots,c_m$ are all those $\mathbf{c}_i$ which do not contain **z**. Then the consequent of the above implication implies

$$\forall\mathbf{y}.\ \mathbf{C}^m_1\vee Q\mathbf{z}.\ \mathbf{C}^n_{m+1}\vee(\mathbf{B}[\![do(\mathbf{y},\mathbf{x})]\!])^*_{(\mathbf{y},\mathbf{x})}$$

that is, $\forall\mathbf{y}.\ \mathbf{C}^m_1\vee(Q\mathbf{z}.\ \mathbf{B}[\![do(\mathbf{y},\mathbf{x})]\!])^*_{(\mathbf{y},\mathbf{x})}$ by definition of *

that is, $\forall\mathbf{y}.\ \mathbf{C}^m_1\vee(\mathbf{A}[\![do(\mathbf{y},\mathbf{x})]\!])^*_{(\mathbf{y},\mathbf{x})}$

We have proved

$$K\vdash\mathbf{A}[\![\mathbf{x}]\!]\supset.\ \forall\mathbf{y}\,C^m_1\vee(\mathbf{A}[\![do(\mathbf{y},\mathbf{x})]\!])^*_{(\mathbf{y},\mathbf{x})}$$

where $\mathbf{c}_1, \ldots, \mathbf{c}_m$ are all those T-terms which occur crucially and are free in **A**.

QED

Another apparent weakness of (→3) is that it allows us to infer that *facts* do not change, but not directly that values of functions do not change. However, (→3) has the further consequence:

Lemma 2

Let $\mathbf{a}=\mathbf{a}[\![\mathbf{x}]\!]$ be a T-term of K, and let $\mathbf{a}_1, \ldots, \mathbf{a}_n$ be all the T-terms, other than **a**, occurring crucially in **a**. Then

$$\forall \mathbf{y} \,.\, (\mathbf{a}\rightarrow \mathbf{y}, \mathbf{x}) \vee (\mathbf{a}_1\rightarrow \mathbf{y}, \mathbf{x}) \vee \ldots \vee (\mathbf{a}_n\rightarrow \mathbf{y}, \mathbf{x}) \vee (\mathbf{a}=\mathbf{a}[\![do(\mathbf{y}, \mathbf{x})]\!])$$

is a theorem of K.

Proof. Let **z** be a variable which does not occur in **a**.

Then an instance of (→3) is:

$$(\mathbf{a}=\mathbf{z}) \supset \,.\, \forall \mathbf{y} \,.\, (\mathbf{a}\rightarrow \mathbf{y}, \mathbf{x}) \vee (\mathbf{a}_1\rightarrow \mathbf{y}, \mathbf{x}) \vee \ldots \vee (\mathbf{a}_n\rightarrow \mathbf{y}, \mathbf{x}) \vee (\mathbf{a}[\![do(\mathbf{y}, \mathbf{x})]\!]=\mathbf{z})$$

and the instance of this obtained by substituting **a** for **z** yields the result.

QED

Making definitions

A certain amount of care is necessary when introducing new non-logical symbols into a kinematic theory. The difficulty is that the schema (→3) will apply to statements involving the new symbols. Thus in extending the vocabulary of a kinematic theory one inevitably has to allow new axioms to be produced. For instance, consider the following set of axioms:

$$P(f(a), b) \tag{1}$$

$$\sim P(f(a), do(c, b)) \tag{2}$$

$$\sim (a\rightarrow c, b) \tag{3}$$

These are consistent. But if we add

$$Q(x) \equiv P(f(x)) \tag{4}$$

where Q is a new predicate letter, then we can easily derive a contradiction:

$$Q(a, b) \quad \text{from (1) and (4)} \tag{5}$$

$$Q(a, do(c, b)) \quad \text{from (5), (3) and (→3)} \tag{6}$$

$$\sim Q(a, do(c, b)) \quad \text{from (2) and (4)} \tag{7}$$

Clearly, the contradiction arises because the definition suppresses a term upon which the value of the definiens depends.

In ordinary first-order logic we have the theorem that if a theory T is consistent, it remains so when we introduce a new predicate symbol **q** and the axiom $\mathbf{q}(x_1, \ldots, x_n)=\mathbf{A}$, where **A** is some statement not containing **q**. In a kinematic theory this theorem is still true provided that **A** contains no expressions with range {*situation*}. In the general case we have a weaker theorem:

Lemma 3

If a kinematic theory is consistent then so is the theory obtained by adding a new predicate symbol $\mathbf{q}$ and the axiom

$$\exists \mathbf{y}(P_{action}(\mathbf{y}) \& ((\mathbf{a}_1 \rightarrow \mathbf{y}, \mathbf{x}) \vee \ldots \vee (\mathbf{a}_n \rightarrow \mathbf{y}, \mathbf{x}))) \vee . \quad \mathbf{q}(\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{x}) \equiv \mathbf{A}[\![\mathbf{x}]\!]$$

where $\mathbf{a}_i$ are all the terms which occur crucially in $\mathbf{A}$ but not in $\mathbf{q}(\mathbf{x}_1, \ldots, \mathbf{x}_m, \mathbf{x})$.

We shall omit the proof, which involves checking that no contradiction can be derived from this axiom using $(\rightarrow 3)$.

I have declared that the three axioms displayed above are consistent. However, to substantiate this it will be necessary to define an appropriate notion of semantics for kinematic theories. This will now be done.

Configurations and scenarios

Let T be a static theory and $K_\Theta(T) = K$ a kinematic extension of it. We shall assume the notational conventions already used above. A model $\mathscr{A}$ of T is called a *configuration*. It represents an actual static state of affairs. K is designed to reason about just such arrangements and how actions change one into another. It would be natural for expressions with range $\{situation\}$ to denote configurations, and expressions with range $\{action\}$ to denote functions from configurations to configurations. We shall define structures for K which have these properties.

A *scenario* $\mathscr{A}$ for $K_\Theta(T)$ is

1. A nonempty set $|\mathscr{A}_s|$ for each sort s of T.

We define $|\mathscr{A}| = \cup |\mathscr{A}_s|$ to be the *universe* of $\mathscr{A}$.

2. For each nonlogical symbol $\mathbf{h}$ in $\Psi - \Theta$, a function or predicate $\mathbf{h}_\mathscr{A}$ of the appropriate degree in $|\mathscr{A}|$.

3. A set $S_\mathscr{A}$ of interpretations of T. Each $s \in S_\mathscr{A}$ has universe $|\mathscr{A}|$, and the interpretation of $\mathbf{h} \in \Psi - \Theta$ is $\mathbf{h}_\mathscr{A}$ specified above. Let $\mathbf{h}^s_\mathscr{A}$ be the function or predicate denoted by $\mathbf{h}$ in $s \in S_\mathscr{A}$.

4. A set $A_\mathscr{A}$ of functions from $S_\mathscr{A}$ to $S_\mathscr{A}$.

5. A function $\Delta_\mathscr{A}$ from $A_\mathscr{A} \times S_\mathscr{A}$ into subsets of $|\mathscr{A}|$.

6. For each nonlogical symbol in Φ, a function or predicate of the appropriate degree in $|\mathscr{A}| \cup A_\mathscr{A} \cup S_\mathscr{A}$, where $A_\mathscr{A}$ contains all denotata of *action* expressions.

7. For the logical predicates $=$ and P_s, the obvious predicates on $|\mathscr{A}| \cup A_\mathscr{A} \cup S_\mathscr{A}$.

We have not specified any special interpretation for nonlogical symbols in Θ_s, $\rightarrow$, or *do*. However,

8. *do* denotes the function which takes $a \in A_\mathscr{A}$ and $s \in S_\mathscr{A}$ into $a(s) \in S_\mathscr{A}$.

9. If $\mathbf{h} \in \Theta^s$, then $\mathbf{h}$ denotes the function which takes $x_1, \ldots, x_n \in |\mathscr{A}|$ and $s \in S_\mathscr{A}$ into $\mathbf{h}^s_\mathscr{A}(x_1, \ldots, x_n)$.

10. $\rightarrow$denotes that predicate which is true of $\langle x_1, x_2, x_3 \rangle$ when

(a) $\rightarrow^{x_3}_{\mathscr{A}}(x_1, x_2)$ is true and $x_2 \in |\mathscr{A}|$

or

(b) $x_1 \in \Delta_{\mathscr{A}}(x_2, x_3)$ and $x_2 \in A_{\mathscr{A}}$.

(Notice here that $\rightarrow$ in (a) is a symbol of T rather than of K.)

It should be clear that a scenario has all the structure which one would expect in an interpretation of K. The function $\Delta_{\mathscr{A}}$ is intended to be the link between objects (members of $|\mathscr{A}|$) and actions: $x \in \Delta_{\mathscr{A}}(a, s)$ when x is causally connected to the action a in situation s, so that performing the action is liable to change some property of x.

A scenario defines the truth-values of ground atoms in the obvious way. We shall assume that this truth definition is extended to all statements of K by the usual recursion, where expressions with range $\{situation\}$ denote members of S, and those with range $\{action\}$ denote members of A. We will also take over the ideas of a *model*, and of a formula being *valid* in a scenario.

A kinematic theory K is of course an ordinary first-order theory and therefore the usual first-order model theory applies to it. The question immediately arises as to the relationship between first-order interpretations of K and scenarios of K.

Lemma 4

There is a $1-1$ truth-preserving correspondence between scenarios of a kinematic theory K and first-order interpretations of K. (Remark: if this seems intuitively obvious to the reader, good. Otherwise, the proof is intended to persuade rather than convince. A full proof would be very long and tedious.)

The notational conventions used earlier will be assumed.

We will give five necessary and sufficient conditions for a first-order structure $\mathscr{A}$ and a scenario $\mathscr{B}$ to be in correspondence.

Condition 1. $|\mathscr{A}_s| = |\mathscr{B}_s|$ for each $s \in S_T$.

Now let $x \in |\mathscr{A}_{situation}|$. We can define a structure $\mathscr{C}_x$ for T as follows. For each sort s of T, we have the set $|\mathscr{A}_s|$. If $\mathbf{h}$ is a nonlogical symbol of T, then

$$\mathbf{h}_{\mathscr{C}_x}(x_1, \ldots, x_n) = \tilde{\mathbf{h}}_{\mathscr{A}}(x_1, \ldots, x_n, x) \text{ if } \mathbf{h} \in \Theta$$
$$= \mathbf{h}_{\mathscr{A}}(x_1, \ldots, x_n) \text{ otherwise}$$

(It follows directly from the definition of kinematic extension that the $\mathbf{h}_{\mathscr{C}_x}$ are defined at the proper places.) For the logical symbols $=$ and P_s, we have the obvious predicates on $\cup|\mathscr{A}_s|$: clearly these are the restrictions to the smaller universe of the predicates denoted by these symbols in $\mathscr{A}$.

Condition 2. $S_{\mathscr{B}} = \{\mathscr{C}_x : x \in |\mathscr{A}_{situation}|\}$.

Now let $y \in |\mathscr{A}_{action}|$. We can define a function f_y from $S_{\mathscr{B}}$ to $S_{\mathscr{B}}$ by

$$f_y(\mathscr{C}_x) = \mathscr{C}_{do_{\mathscr{A}}(y,x)}.$$

Condition 3. $A_{\mathscr{B}} = \{f_y : y \in |\mathscr{A}_{action}|\}$.

Define a function $\Delta_{\mathscr{A}}$ from $|\mathscr{A}_{situation}| \times |\mathscr{A}_{action}|$ to subsets of $|\mathscr{A}|$ by: $x \in \Delta_{\mathscr{A}}(y, z)$ iff $\rightarrow_{\mathscr{A}}(x, y, z)$ is true.

Condition 4. $\Delta_{\mathscr{A}} = \Delta_{\mathscr{B}}$.

Condition 5. The mappings $x \leftrightarrow \mathscr{C}_x$, $y \leftrightarrow f_y$, defined above extend to an isomorphism between the interpretations $\mathbf{h}_{\mathscr{A}}$ and $\mathbf{h}_{\mathscr{B}}$ of symbols in Φ.

Clearly this correspondence is $1-1$ up to isomorphism, and preserves the truth-values of ground atoms.

QED

This shows that the completeness theorem can be transferred immediately to the intuitive semantics.

It remains to show that the three causality axioms do indeed capture the intuitions outlined in the introduction. We shall require one more lemma before the main result can be stated.

Lemma 5*

A first-order theory (with the morphology of a kinematic theory) contains all instances of the schema $(\rightarrow 3)$ iff it contains all instances of the two schemas

$$(\mathbf{x}_1 \rightarrow \mathbf{y}, \mathbf{x}) \vee \ldots \vee (\mathbf{x}_n \rightarrow \mathbf{y}, \mathbf{x}) \vee . \\ \mathbf{p}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}) \equiv \mathbf{p}(\mathbf{x}_1, \ldots, \mathbf{x}_n, do(\mathbf{y}, \mathbf{x})) \quad (\rightarrow 4)$$

$$(\mathbf{x}_1 \rightarrow \mathbf{y}, \mathbf{x}) \vee \ldots \vee (\mathbf{x}_n \rightarrow \mathbf{y}, \mathbf{x}) \vee (\mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}) \rightarrow \mathbf{y}, \mathbf{x}) \vee \\ \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{x}) = \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_n, do(\mathbf{y}, \mathbf{x})) \quad (\rightarrow 5)$$

The proof of 'only if' is trivial since $(\rightarrow 4)$ follows directly from two instances of $(\rightarrow 3)$, and $(\rightarrow 5)$ follows directly from lemma 2.

Now suppose a theory contains all instances of $(\rightarrow 5)$. We shall show by induction on the structure of the term $\mathbf{a}$ that the theory contains all instances of the schema

$$(\mathbf{a} \rightarrow \mathbf{y}, \mathbf{x}) \vee (\mathbf{a}_1 \rightarrow \mathbf{y}, \mathbf{x}) \vee \ldots \vee (\mathbf{a}_n \rightarrow \mathbf{y}, \mathbf{x}) \vee (\mathbf{a} = \mathbf{a}[\![do(\mathbf{y}, \mathbf{x})]\!]) \quad (\rightarrow 6)$$

where $\mathbf{a} = \mathbf{a}[\![\mathbf{x}]\!]$ is a T-term and the $\mathbf{a}_i$ are all the terms other than $\mathbf{a}$ occurring crucially in $\mathbf{a}$ (cf. lemma 2).

If $\mathbf{a}$ contains just one function symbol then $(\rightarrow 6)$ is $(\rightarrow 5)$, by definition of 'crucially'. Now suppose $\mathbf{a}$ is $\mathbf{f}(\mathbf{a}_1, \ldots, \mathbf{a}_m, \mathbf{x})$, then the terms occurring crucially in $\mathbf{a}$ are $\mathbf{a}$ itself, the $\mathbf{a}_i$, and every term which occurs crucially in some $\mathbf{a}_i$. By induction hypothesis, $(\rightarrow 6)$ holds for each $\mathbf{a}_i$. Moreover, by $(\rightarrow 5)$, we have

$$(\mathbf{a}_1 \rightarrow \mathbf{y}, \mathbf{x}) \vee \ldots \vee (\mathbf{a}_m \rightarrow \mathbf{y}, \mathbf{x}) \vee (\mathbf{a} \rightarrow \mathbf{y}, \mathbf{x}) \vee \mathbf{a} = \mathbf{f}(\mathbf{a}_1, \ldots, \mathbf{a}_m, do(\mathbf{y}, \mathbf{x})).$$

The substitutivity of equality then yields $(\rightarrow 6)$ directly. Now suppose $\mathbf{a}$ is $\mathbf{f}(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ where $\mathbf{a}_m$ is not $\mathbf{x}$. Then no other $\mathbf{a}_i$ is $\mathbf{x}$ (by the morphological rules for a kinematic theory) and so the terms occurring crucially in $\mathbf{a}$ are just those occurring crucially in some $\mathbf{a}_i$. Using the induction hypothesis and the substitutivity of equality gives $(\rightarrow 6)$ directly.

Now, suppose a theory contains $(\rightarrow 4)$ and $(\rightarrow 6)$. Let $\mathbf{A}[\![\mathbf{x}]\!]$ be a literal. If $\mathbf{A} = \mathbf{p}(\mathbf{a}_1, \ldots, \mathbf{a}_m, \mathbf{x})$ or $\sim\mathbf{p}(\mathbf{a}_1, \ldots, \mathbf{a}_m, \mathbf{x})$ then the terms occurring crucially

* This lemma was suggested by a remark of Gordon Plotkin.

in A are the $\mathbf{a}_i$ and these terms which occur crucially in some $\mathbf{a}_i$. Using the following instance of (→4),

$$(\mathbf{a}_1 \rightarrow \mathbf{y}, \mathbf{x}) \vee \ldots \vee (\mathbf{a}_m \rightarrow \mathbf{y}, \mathbf{x}) \vee . \; \mathbf{p}(\mathbf{a}_1, \ldots, \mathbf{a}_m, \mathbf{x}) \equiv \mathbf{p}(\mathbf{a}_1, \ldots, \mathbf{a}_m, do(\mathbf{y}, \mathbf{x})),$$

the appropriate instances of (→6) for each $\mathbf{a}_i$, and the substitutivity of equality, gives (→3). If $\mathbf{A} = \mathbf{p}(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ where $\mathbf{a}_m$ is not $\mathbf{x}$, we need only use (→6) and equality, as in the second case above.

QED

Now the main result can be proved.

Theorem

(1) The axiom (→1) is valid in a scenario $\mathscr{A}$ iff $\rightarrow^x_{\mathscr{A}}$ is reflexive for every $x \in S_{\mathscr{A}}$.

(2) The axiom (→2) is valid in a scenario $\mathscr{A}$ iff $\rightarrow^x_{\mathscr{A}}$ is transitive for every $x \in S_{\mathscr{A}}$, and in addition

$$x \in \Delta_{\mathscr{A}}(y, s) \text{ and } \rightarrow^s_{\mathscr{A}}(z, x) \text{ together imply } z \in \Delta_{\mathscr{A}}(y, s).$$

(3) The axiom schema (→3) is valid in a scenario $\mathscr{A}$ iff $\mathscr{A}$ obeys the *causality condition*:

$$\text{If } x_1, \ldots, x_n, \mathbf{h}^s_{\mathscr{A}}(x_1, \ldots, x_n) \notin \Delta_{\mathscr{A}}(a, s)$$
$$\text{then } \mathbf{h}^s_{\mathscr{A}}(x_1, \ldots, x_n) = \mathbf{h}^{a(s)}_{\mathscr{A}}(x_1, \ldots, x_n)$$

for every $\mathbf{h} \in \Theta$.

Proof. (1) Trivial. (2) Easy. (3) The schema (→3) is valid in $\mathscr{A}$ iff both of (→4) and (→5) are, by lemmas 4 and 5. Now suppose $\mathscr{A}$ satisfies the causality condition for each $\mathbf{h} \in \Theta$, and suppose (→4) is not valid in $\mathscr{A}$. Then for some predicate symbol $\mathbf{p}$, we have $((\mathbf{i}_1 \rightarrow \mathbf{j}, \mathbf{i}) \vee \ldots \vee (\mathbf{i}_n \rightarrow \mathbf{j}, \mathbf{i}) \vee . \; \mathbf{p}(\mathbf{i}_1, \ldots, \mathbf{i}_n, \mathbf{i}) \equiv \mathbf{p}(\mathbf{i}_1, \ldots, \mathbf{i}_n, do(\mathbf{j}, \mathbf{i})))_{\mathscr{A}} = false$. I.e., where $\mathbf{i}$ is the name of i, etc., we have $i_1, \ldots, i_n \notin \Delta_{\mathscr{A}}(j, i)$ and $\mathbf{p}^i_{\mathscr{A}}(i_1, \ldots, i_n) \neq \mathbf{p}^{j(i)}_{\mathscr{A}}(i_1, \ldots, i_n)$. Clearly $\mathbf{p}^i_{\mathscr{A}}(i_1, \ldots, i_n) \notin \Delta_{\mathscr{A}}(j, i)$ and thus $\mathbf{p}$ violates the causality condition. The argument when (→5) is not valid in $\mathscr{A}$ is exactly similar.

Now suppose that both of (→4), (→5) are valid in $\mathscr{A}$. Then we can show directly, using the same sort of construction as above and remembering that every individual in $\mathscr{A}$ has a name in the language, that the causality condition is satisfied for each $\mathbf{h} \in \Theta$.

QED

We can now display a model of the axioms mentioned earlier:

(1) $P(f(a), b)$ (2) $\sim P(f(a), do(c, b))$ (3) $\sim(a \rightarrow c, b)$

Let c_1 and c_2 be two structures with universe $\{a, d\}$ in which → denotes identity and $f(a) = f(d) = d$. In c_1, P is true of d and false of a; in c_2, P is everywhere false. Let $S_{\mathscr{A}}$ be $\{c_1, c_2\}$. Let $A_{\mathscr{A}}$ be $\{c\}$ where $c(c_1) = c_2$ and $c(c_2) = c_2$. Let $\Delta_{\mathscr{A}}(c, c_1) = \Delta_{\mathscr{A}}(c, c_2) = \{a\}$. Let P, f, a and c denote themselves and b denote c_1. Then $\mathscr{A}$ is a scenario which is a model of (1)–(3) and (→1)–(→3).

The causality condition can be seen as a specification that the effect of any action should be *local*. It guarantees that the configurations before and after an action are isomorphic, except on the individuals singled out by Δ.

AN EXAMPLE: THE MONKEY AND THE BANANAS

In this section I shall give a set of axioms which partly describe a simple world such as might be inhabited by a primitive monkey. [The toy world described in this section is based upon Robin Popplestone's axioms for Freddy, the Edinburgh robot (Popplestone 1970).] The axioms are not complete and their shortcomings will be discussed, but they are sufficient to prove that the monkey in the classical puzzle can get his bananas.

There are four sorts in the static language: *thing*, *place*, *integer* and *monkey*. The nonlogical symbols in the static language are: HELD and PLATFORM, predicates on *things*; M, a constant of sort *monkey*; PLACE, a function from {*thing*, *monkey*} to *places*; ON and UNDER, functions from *things* to *places*; HEIGHT and LOCATION, functions from *places* to *integers*; HT, a function from *things* to *integers*, and finally the arithmetic symbols $<$ and $+$. We shall also of course have the binary predicate $\rightarrow$ and it will be convenient to assume that the ranges of the arguments of $\rightarrow$ do not include *place* or *integer*, thus avoiding the axioms $P_{place}(x) \supset .(x \rightarrow y,\ s) \supset x = y$ and $P_{integer}(x) \supset .(x \rightarrow y,\ s) \supset x = y$.

The ideas behind these are as follows. *Things* are ordinary physical objects; *places* are positions in some homomorph of Euclidean 3-space. The functions *height* and *location* define the coordinate system in this space. I envisage a fairly precise vertical distance measure but a rather loose and tolerant horizontal measure (which is, however, coded into integers in some way), so that several different objects can be at the same place. The *place* of an object is the place it is sitting at: the place which is *on* an object is the top of it (for example, the box): the place *under* an object is the patch of floor beneath it. The *ht* of an object is the vertical distance between the *place* of it and *on* of it. An object is a *Platform* when it is firm and large enough for the monkey to stand on it. A thing is *Held* when the monkey has it in his hand.

We shall assume that the static theory contains the following axioms:

P1. $place(x) = on(y) \supset (x \rightarrow y)$

P2. $place(x) \neq on(x)$

P3. $height(x) = 0 \supset \exists y(x = on(y))$

P4. $(M \rightarrow x) \supset .\ x = M \vee place(M) = on(x)$

P5. $location(on(x)) = location(place(x))$
$\& height(on(x)) = height(place(x)) + ht(x)$

P6. $\forall z\,(place(x) \neq on(z)) \supset .\ location(under(x)) = location(place(x))$
$\& height(under(x)) = 0$

P7. $(x \rightarrow y) \supset .\ location(place(x)) = location(place(y))$

P8. $Held(x) \supset .(x \rightarrow y) \equiv y = M$

This is not a complete set of axioms, of course, and I do not have such a set. Their interpretation should be fairly obvious. P3 says that things on the floor are indeed *on* something: P4 says that the only way in which the monkey can be attached to a thing is by being *on* it. The antecedent of P6 stops the monkey inferring anything about places *under* objects. P7 says there is no causality at a distance. P8 says that *Held* objects are attached only to the monkey.

The chief shortcoming of this theory at present is that there is no way of expressing facts of *solidity* or *weight*. These both seem to present formidable problems.

Now we shall extend this theory to a kinematic theory. The set Θ is to be $\{place, on, under, \rightarrow, Held\}$. We add the nonlogical symbols: MOVE, a function from *places* to *actions*, and CLIMB, UNCLIMB, GET and PUT, functions from *things* to *actions*, as well as *do*. For each action we have to supply the appropriate law of motion defining the direct effect of the action, and also specify what objects are directly connected to actions. We shall assume the following axioms:

Move

M1. $((x \rightarrow move\,(y)) \equiv (x \rightarrow M))[\![s]\!]$

M2. $(height\,(place\,(M, s)) = height\,(y) = 0\ \&\ (x \rightarrow M, s))$
$\supset place\,(x, do\,(move\,(y), s)) = y$

M3. $do\,(move\,(place\,(M, s)), s) = s^{\frown}$

M4. $Held\,(x, s) \supset Held\,(x, do\,(move\,(y), s))$

Climb

C1. $((x \rightarrow climb\,(y)) \equiv (x \rightarrow unclimb\,(y)) \equiv (x \rightarrow M))[\![s]\!]$

C2. $(Platform\,(x)\ \&\ ht\,(x) < 3\ \&\ \forall y(\sim Held\,(y, s))) \supset .$
$(place\,(M, s) = place\,(x, s) \supset place\,(M, do\,((climb\,(x), s)) = on\,(x, s))$
$\&\ (place\,(M, s) = on\,(x, s) \supset place\,(M, do\,(unclimb\,(x), s)) = place\,(x, s))$

Get and Put

G1. $((x \rightarrow get\,(y)) \equiv (x \rightarrow put\,(y)) \equiv (x \rightarrow y))[\![s]\!]$

G.2 $(height\,(place\,(x, s)) < height\,(place\,(M, s)) + 3$
$\&\ location\,(place\,(x, s)) = location\,(place\,(M, s))) \supset$
$Held\,(x, do\,(get\,(x), s))$

G3. $Held\,(x, s) \supset (\sim(x \rightarrow M)\ \&\ place\,(x) = place\,(M))[\![do\,(put\,(x), s)]\!]$

Causality

Ca. $(x \rightarrow y, s) \supset .\,(z \rightarrow x, do\,(y, s)) \supset (z \rightarrow x, s)$

These are all fairly obvious except perhaps for Ca. M2 lets the monkey move to any place along the floor, and take things with him. M3 says that doing nothing really is doing nothing. M4 says the monkey does not drop things. C2 lets the monkey climb onto and off platforms which are not too high, provided his hands are free. G2 lets him get hold of things which are not too

far above him. G3 lets him put things down at his feet (as opposed to throwing them away). The special causality axiom is extremely useful: it says that the objects affected by an action do not themselves pick up other objects as the action is performed. This can be done, and the axiom thereby falsified, by sliding an object underneath another while supporting the latter in some way, as shown for instance in figure 2. (*Ca* might be called the 'no-hooks' axiom). We shall assume the monkey is not so clever.

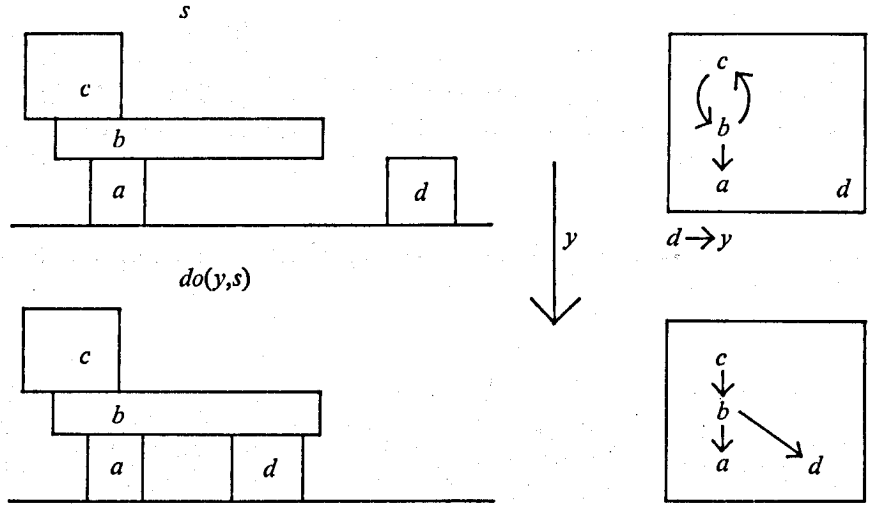

Figure 2

Again, this is obviously not a completed theory. However, we can now set up the puzzle and let the monkey get the bananas. We need some facts about the initial situation s_0, and we shall use the symbols BOX and BA(nanas):

o1. $height\,(place\,(Box, s_0))=height\,(place\,(M, s_0))=0$

o2. $height\,(place\,(Ba, s_0))=3$

o3. $Platform\,(Box)\ \&\ ht\,(Box)=2$

o4. $(x\rightarrow M, s_0)\supset x=M$

o5. $(x\rightarrow Box, s_0)\supset x=Box$

o6. $place\,(Ba, s_0)\neq on\,(x, s_0)$

A proof that the monkey can get the bananas follows. He has to go to the box, *get* it, *move* under the bananas, *put* the box, *climb* on it, and *get* the bananas. Let $\mathbf{a}=move\,(place\,(Box, s_0))$, $\mathbf{s}_1=do\,(\mathbf{a}, s_0)$, $\mathbf{b}=get\,(Box)$, $\mathbf{s}_2=do\,(\mathbf{b}, \mathbf{s}_1)$.

1. $place\,(M, \mathbf{s}_1)=place\,(Box, s_0)$ — from o1., →1, M2.
2. $(x\rightarrow \mathbf{a}, s_0)\equiv x=M$ — from M1., o4, →1.

Using 2., we can infer

3. $place\,(Box, \mathbf{s}_1)=place\,(Box, s_0)$ — from 2., Lemma 2.

4. $(x \rightarrow Box, \mathbf{s}_1) \supset . x = Box \vee x = M$ — from 2., O5, →3.
5. $place(M, \mathbf{s}_1) = place(Box, \mathbf{s}_1) \neq on(Box, \mathbf{s}_1)$ — from 1., 3., P2.
6. $(x \rightarrow Box, \mathbf{s}_1) \supset x = Box$ — from 4., 5., P4.
7. $Held(Box, \mathbf{s}_2)$ — from 5., G2.
8. $(x \rightarrow \mathbf{b}, \mathbf{s}_1) \equiv x = Box$ — from 6., G1, →1.
9. $place(M, \mathbf{s}_2) = place(M, \mathbf{s}_1)$ — from 8., Lemma 2.
10. $height(place(M, \mathbf{s}_2)) = 0$ — from 9., 1., O1.

Now, the monkey can move under the bananas, since

11. $height(under(Ba, s_0)) = 0$ & $location(under(Ba, s_0)) = location(place(Ba, s_0))$ — from O6, P6..

Thus, let $\mathbf{c} = move(under(Ba, s_0))$, $\mathbf{s}_3 = do(\mathbf{c}, \mathbf{s}_2)$.

12. $place(M, \mathbf{s}_3) = under(Ba, s_0)$ — from 10; 11; →1; M2.
13. $place(Box, \mathbf{s}_3) = under(Ba, s_0)$ — from 10., 11., 7., P8.; M2.
14. $(x \rightarrow \mathbf{c}, \mathbf{s}_2) \equiv (x \rightarrow M, \mathbf{s}_2)$ — from M1.

In order to make 14. useful we have to find out what is attached to the monkey.

15. $((x \rightarrow M, \mathbf{s}_2) \supset . (x \rightarrow M, \mathbf{s}_1) \vee x = Box)$ — from 8., →3.
    & $(x \rightarrow M, \mathbf{s}_1) \supset . x = M$ — from 2., O4., Ca.
16. $(x \rightarrow \mathbf{c}, \mathbf{s}_2) \equiv . x = M \vee x = Box$ — from 14., 15., →1.
17. $Held(Box, \mathbf{s}_3)$ — from 7., M4.

Now, the monkey must drop the box before he can climb onto it.
Let $\mathbf{d} = put(Box)$, $\mathbf{s}_3 = do(\mathbf{d}, \mathbf{s}_3)$.

18. $\sim(Box \rightarrow M, \mathbf{s}_4)$ & $place(Box, \mathbf{s}_4) = place(M, \mathbf{s}_4)$ — from 17., G3., 13.
19. $(x \rightarrow \mathbf{d}, \mathbf{s}_3) \equiv (x \rightarrow Box, \mathbf{s}_3)$ — from G1.

In order to make 19. useful we must find out what is attached to the box.

20. $(x \rightarrow Box, \mathbf{s}_3) \supset (x \rightarrow Box, \mathbf{s}_2)$ — from 16., Ca.
    & $(x \rightarrow Box, \mathbf{s}_2) \supset (x \rightarrow Box, \mathbf{s}_1)$ — from 8., Ca.

hence

21. $(x \rightarrow Box, \mathbf{s}_3) \equiv x = Box$ — from 20., 6., →1.
22. $(x \rightarrow \mathbf{d}, \mathbf{s}_3) \equiv x = Box$ — from 19., 21.

Now we can infer that the monkey has not moved:

23. $place(M, \mathbf{s}_4) = place(M, \mathbf{s}_3) = under(Ba, s_0)$ — from 22., Lemma 2, 12.

and hence

24. $place(Box, \mathbf{s}_4) = under(Ba, s_0)$ — from 23., 18.

Now we need to show that the monkey's hands are free. This is fairly direct, but we shall do it in detail for clarity:

25. $(x \rightarrow M, \mathbf{s}_2) \supset . x = M \vee x = Box$ — from 15.
26. $(x \rightarrow M, \mathbf{s}_3) \supset (x \rightarrow M, \mathbf{s}_2)$ — from 16., Ca.
27. $(x \rightarrow M, \mathbf{s}_3) \supset . x = M \vee x = Box$ — from 25., 26.
28. $((x \rightarrow M, \mathbf{s}_4) \supset (x \rightarrow M, \mathbf{s}_3)) \vee x = Box$ — from 22., →3.

29. $(x \rightarrow M, s_4) \supset . x = M \vee x = Box$ from 27., 28.
30. $(x \rightarrow M, s_4) \supset x = M$ from 29., 18.
31. $\sim Held(x, s_4)$ from 30., P8.

Now his hands are free, he can climb on the box. Let $e = climb(Box)$, $s_5 = do(e, s_4)$.

32. $place(M, s_5) = on(Box, s_4)$ from O3., 31.; 23., 24.; C2.
33. $(x \rightarrow e, s_4) \equiv x = M$ from C1., 31., $\rightarrow$1.

Now we can infer that the bananas are still where they were:

34. $place(Ba, s_5) = place(Ba, s_0)$ from 2., 8., 16., 22., 33., Lemma 2 (5 times)

Now tracing back, we see that

35. $location(place(M, s_5)) = location(Ba, s_5))$ from 32., P5., 24., 11., 35.

Now we have to find out how high the monkey is, and this is easy:

36. $place(Box, s_5) = place(Box, s_4)$ from 33., Lemma 2.
37. $height(place(Box, s_5)) = 0$ from 36., 24., 11.
38. $height(place(M, s_5)) = 2$ from 37., 32., P5., O3.
39. $height(place(Ba, s_5)) = 3$ from 34., O2.

And thus, finally:

40. $Held(Ba, do(get(Ba), s_5))$ from 35., 38., 39., G2.

The reader might like to try proving that the monkey can get down off the box, with the bananas. It is not as easy as it looks, since his hands have to be free before he can climb down.

This proof, although long, is fairly natural. Its length is due partly to the fact that I have carried it out in detail at several points. Familiarity with kinematic systems enables one to see immediately such consequences as 31 from 25: clearly one would want to incorporate such 'macro' deductions in some efficient way in a mechanized proof-seeking procedure. I do not claim that kinematic systems as presented here are particularly machine-oriented, only that one can in fact set up puzzles and solve them within the system. They are intended to be *epistemologically*, rather than *heuristically*, adequate (cf. McCarthy and Hayes 1969).

The axioms presented earlier do not contain many 'frame' axioms. The most unattractive such is M4. The need for M4 arises because a thing held is attached to the monkey and therefore to the action *move* (z). Thus nothing can be inferred about its properties when *move* (z) is performed. This illustrates a general weakness of kinematic theories. When some property, however trivial, of an object is altered by an action, then the object must have been connected to the action, by $(\rightarrow 3)$, and therefore *any* property of the object is liable to have changed when the action is performed. This is the *local* version of the frame problem. The *global* problem is handled by $(\rightarrow 3)$. Now, some obvious progress can be made here by re-introducing the notion of a frame, as will be remarked later. But it is still a difficulty, and will become more serious as axiom systems become richer. Thus suppose we had predicates

of colour: then we would be obliged to have an axiom saying that objects attached to the monkey do not change their colour during a *move*.

The monkey axioms are capable of extension in several directions. In particular, one can have the monkey building towers of bricks. Here an interesting point emerges. It turns out to be a useful heuristic to pay attention most to those objects to which few other objects are causally related. For instance, if one wants a brick, take it off the top of a tower rather than the bottom.

One of the most useful aspects of developing axiom systems within kinematic theories is the way in which such heuristic principles seem to emerge. Another one was followed in constructing the above proof: to wit, when a new action is contemplated, first find out exactly what objects are connected to it. This information is usually rapidly obtainable (at least in this simple world) and almost certain to be useful subsequently.

DISCUSSION

Presuppositions and non-deductive reasoning

The reader will no doubt have been adversely impressed by the length of the proof that the monkey can get the bananas. This contrasts quite dramatically with the proofs obtained by, for example, Cordell Green from his axiomatizations of the puzzle. To some extent this is due to the lack of special 'banana' axioms, or similar tricks, in the present formulation. However there does arise in a kinematic system the necessity of continually proving that objects are *not* connected to one another or to actions. Such subproofs occupy a considerable amount of time for our monkey.

Now, it may be argued that this is unintuitive and that a better system could be got by arranging that such facts were somehow assumed to be true unless there were explicit statements to the contrary. Introspection seems to show that humans do not continually find the need to prove that objects are *not* connected together. For instance, if one presents the usual verbal account to someone, and he suggests the obvious solution, and you then tell him that it will not work because the string from which the bananas are hanging passes over two pulleys and is connected to the box, so that when the monkey moves the box under the bananas they ascend out of his reach; then he will probably object that you should have told him that in the first place. He will not be impressed by your arguments that you have not said anything which is contradicted by the statement of the problem, but only added a little to the description of the world. He will have assumed that such complexity is not present since it was not mentioned. Is there a logic which embodies assumptions like this? The answer, unfortunately, is no: assuming, that is, that 'logic' means a deductive system. Clearly, so long as it is consistent to add statements asserting that objects are connected, it is inconsistent to infer, without adequate grounds, that they are not so connected. Thus, let S be a formal transcription of an intuitive formulation of the monkey and bananas

puzzle. Then in such a logic, a statement C asserting that (say) the box is not connected to the bananas follows from S: $S \vdash C$. But, the negation of C is supposed to be consistent with S.

There are two remarks which can be made. First, if the monkey has eyes, then he can *see* that objects are not connected after the performance of actions. Secondly, one can envisage the construction of a non-deductive logical system based upon a kinematic system K which would operate by making such assumptions, that is, by strengthening instances of $(\rightarrow 3)$ by deleting some of their antecedents. Such a system would have shorter proofs than K, could be more intuitive in its reasoning than K, and would, of course, be inconsistent. However, the inconsistency would be of a rather mild kind and hopefully comparatively easy to control. It should not be too difficult, when the contradiction arises, to identify the particular presupposition which engendered it. [Gerald Sussman has pointed out to me that *Planner*, the robot deduction engine developed at MIT by Carl Hewitt (1969), provides facilities for just this kind of procedure.]

Modal logic and plans

Clearly, scenarios bear close resemblances to the standard Kripke semantics for modal logics. Situations are possible worlds; each action defines an alternativeness relation; *do* enables us to write modal-ish statements. The reader may wonder why none of the notorious difficulties of referential opacity arise in kinematic systems. The reason is that we assume that individuals are not created or destroyed by actions. Thus every member of the set of situations has the same universe. I have discussed elsewhere some circumstances in which one might want to relax such an assumption (Hayes 1970), but there are others more pressing.

It might be thought that most objects in a robot's environment are likely to be fairly permanent, and I would agree with this provided it is understood to refer to physical objects. However it is more plausible that actions may come into and go out of existence as time passes. If we allowed this to happen, the logic would become considerably more complicated and special rules about quantification over actions would have to be constructed. The utility of this would be that the robot would be able to assert that in some future situation an action of a required kind will exist. Such a statement can be regarded as a statement of confidence in the robot's own abilities to think of a plan. Such statements seem to play an essential role in the formulation, within first-order logic, of GPS-like means-end analysis.

This idea can be taken a stage further by allowing a special class of actions called intellectual actions. Then an assertion of the kind mentioned above might be phrased in the following way:

$$\mathbf{u}[\![s]\!] \supset \exists x \,.\, \mathbf{v}[x] \;\&\; x \in do\,(subplan, s)$$

where *subplan* is an intellectual action and $\in$ denotes existence in a situation.

It seems to me that this provides a natural way of incorporating means-end analysis into logical plan-making. It provides yet another reason for wanting an adequate mechanization of first-order modal logic.

Frames

The notion of a frame can be carried over with advantage to kinematic systems. Recall that a frame is a classification of the nonlogical symbols of a theory so that changes to the values of symbols in one part of the classification do not affect the values of symbols in the other parts. Thus predicates of colour and predicates of location might be classified apart in a frame. To extend the idea to a kinematic system we have only to provide a causality relation for each block of the classification. The axioms $(\rightarrow 1)$ and $(\rightarrow 2)$ are stated for each of these relations separately, and in the statement of $(\rightarrow 3)$ we need only mention terms and atoms whose main symbol is in the same part of the classification as the particular causality relation being used.

The use of a device like this would have greatly simplified the monkey and bananas proof. The frame would have had two blocks, one containing statements of position and the other containing statements of connection.

A more flexible approach to the same goal would be to add an extra argument place to the causality relation, and a new sort to the theory, called *mode*. The new argument place is to accept only terms with range $\{mode\}$. There should be enough modes to distinguish the various blocks of the frame, but we now have the ability to state general laws of causality and also to state laws which move across frame boundaries in potentially complicated ways. For instance, the space of modes might have any structure, for example, that of a semigroup.

Theory of computation

It is of course essential to provide eventually some means of constructing plans with loops, or some device of equal power. I do not foresee any major problems in grafting on to the present system some theory of algorithms, but it does seem that the constraints which the causality axioms impose upon plan formation will result in a comparatively weak theory. That is, it will be rather more difficult than usual to prove desirable properties of programs such as termination and equivalence. In the case of a simple loop, for instance, the induction hypothesis for the proof of termination must ensure that the preconditions of freedom for the nth iteration have not been violated by what was done during the $(n-1)$th iteration.

It seems useful to regard the present theory as an initial attempt at a theory of *data structures* which model the real world, as opposed to a theory of *algorithms* operating on these data structures.

The natural logical language in which to embed a theory of algorithms is 3-valued, rather than the classical 2-valued, predicate calculus. The use of this, in fact, would have yielded a more natural system even at the present

level of complexity. For instance, the statement of the necessary conditions for the feasibility of an action, and the statement of the effect of doing the action, could be separated. At present it is necessary to merge them into a single implication (as in the first *move* axiom in the monkey and bananas problem). In the semantics, it would no longer be necessary to insist that every action was a total function on the set of situations, a requirement which may have already struck the reader as unnatural.

Naturalness

In spite of its length, I would claim that every step in the monkey and bananas proof is intuitively convincing. If one had to convince an intellectual moron that the monkey could get the bananas, one might use similar arguments to counter objections he might raise about things moving spontaneously. Such principles as there being no action at a distance, and the rather unusual *ca3*, seem to correspond to our fundamental intuitions about the physical world. Of course, both Freddy's and the monkey's worlds are highly simplified. It is interesting, and quite difficult, to invent principles of causation which are valid in a wider context and are as useful as, for instance, *ca3*.

Any solution of the frame problem will involve some analysis of the idea of causality. The fact that it *is* a problem at all reflects what Simon has called the 'empty world hypothesis' (Simon 1967). For, if the real world was highly interconnected so that small changes in one place invariably led to drastic alterations in large areas, then it would not be surprising that one could not infer anything about one situation given some facts about another. The sense of frustration one experiences in attempting to prove simple strategies reflects our intuitive knowledge that the world is a fairly quiet place. One does not, therefore, expect that a purely 'logical' device will yield an adequate solution since the problem is essentially a 'physical' one. The aim of this paper has been to initiate a study of logical theories of physical causality.

MECHANIZATION

No conventional theorem-proving program could hope to find the monkey and bananas proof from the axioms without some very sophisticated heuristics, even supposing it had some method of handling the axiom schema ($\rightarrow 3$). Gordon Plotkin (personal communication) has devised such a method which fits into the usual resolution format for first-order logic. This is possible because ($\rightarrow 3$) applies to literals, rather than more complicated formulae. But such a restriction makes the proof even longer, since we are unable to use lemma 1 and lemma 2 directly, but have, in effect, to prove them over again each time they are to be used.

Some progress might be made within the conventional framework by inventing clever heuristics. Thus it seems a good idea to determine as soon as possible what objects are attached to actions, and subsequently delay

using (→3) until it is necessary to prove that some property is unchanged. However even to recognize such a need is a fairly sophisticated matter, presupposing some kind of goal-oriented search.

It seems to me that real progress will be made only by constructing a more rigid control structure from within which to conduct a search for a proof. The kind of supervisor sketched in McCarthy (1959) is an example. It seems reasonable that the axioms for such a special notion as causality would need special treatment. One would not expect a general treatment suitable for any axiom system, such as heuristic search, to take adequate account of such an axiom set.

Such a control structure would also facilitate the use of 'macro' inferences; that is, standard small subproofs with few parameters that need to be re-used many times. There is a clear need for such macros in conducting such proofs as the one for the monkey and bananas. However, the use of macros can, unless very carefully controlled, give rise to a combinatorial explosion of its own peculiar kind, especially if it is relatively easy to produce new macros. It is very difficult to control macro generation and expansion from within the simple heuristic search mechanism.

It will be quite difficult to produce such a theorem-proving executive. For the present, the more pressing problems seem to be improving and extending the logic and gaining experience with particular axiomatizations. Any practical implementation must wait upon an adequate theoretical foundation.

Acknowledgements

This work was supported by the Science Research Council. I would like to thank Bernard Meltzer for his encouragement and, together with Bruce Anderson, Rod Burstall, Bob Kowalski, Donald Kuehner, Robin Milner, Gordon Plotkin, Robin Popplestone and Gerald Sussman, for useful conversations and criticisms; and Robin Popplestone for inventing his Freddy axioms. Particular thanks are due to Gordon Plotkin for finding two major errors in an earlier draft.

REFERENCES

Fikes, R.E. (1970) Ref-Arf: A System for solving problems stated as procedures. *Art. Int.*, **1**, 27–120.

Foster, J.M. & Elcock, E.W. (1969) ABSYS 1: An incremental compiler for assertions. *Machine Intelligence 4*, pp. 423–9 (eds Meltzer, B. & Michie, D.). Edinburgh: Edinburgh University Press.

Green, C.C. (1969) Application of theorem proving to problem solving. *Proc. Int. Joint Conf. on Art. Int.*, pp. 219–40. Washington DC.

Hayes, P.J. (1970) Robotologic. *Machine Intelligence 5*, pp. 533–54 (eds Meltzer, B. & Michie, D.). Edinburgh: Edinburgh University Press.

Hewitt, C. (1969) PLANNER: A language for proving theorems in robots. *Proc. Int. Joint Conf. on Art. Int.*, pp. 295–302. Washington DC.

Manna, Z. (1970) The correctness of nondeterministic programs. *Art. Int.*, **1**, 1–26.

McCarthy, J. (1959) Programs with common sense. *Semantic Information Processing*, pp. 403–17 (ed. Minsky, M.). Cambridge, Mass.: MIT Press.

McCarthy, J. & Hayes, P.J. (1969) Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence 4*, pp. 463–502 (eds Meltzer, B. & Michie, D.). Edinburgh: Edinburgh University Press.

Mendelson, E. (1964) *Introduction to Mathematical Logic*. New York: Van Nostrand.

Minsky, M. (1961) Descriptive languages and problem solving. *Semantic Information Processing*, pp. 413–24 (ed. Minsky, M.). Cambridge, Mass.: MIT Press.

Popplestone, R. (1970) Freddy, things and sets. Internal memorandum, Dept of Machine Intelligence and Perception, University of Edinburgh.

Schoenfield, J.R. (1967) *Mathematical Logic*. London: Addison-Wesley.

Simon, H.A. (1967) The logic of heuristic decision making. *The logic of decision and action* (ed. Rescher, N.). Pittsburgh: University of Pittsburgh Press.