LABORATORY FOR
COMPUTER SCIENCE

MASSACHUSETTS
INSTITUTE OF
TECHNOLOGY

MIT/LCS/TM-305

REPRESENTING CHANGE

ELISHA SACKS

MAY 1986

545 TECHNOLOGY SQUARE, CAMBRIDGE, MASSACHUSETTS 02139

Representing Change

Elisha Sacks

MIT Laboratory for Computer Science
545 Technology Square, Room 370
Cambridge, MA 02139
U.S.A.
Tel: (617) 253-3447
Net: ELISHA@MIT-ZERMATT.EDU

Abstract

This paper evaluates knowledge representations for time-dependent information. It compares recent work by Moore, McDermott, and Allen with an earlier proposal by McCarthy and Hayes. Moore's formalism is faulted for its needless and unmotivated complexity and a simpler alternative is outlined. McDermott's formalism is proved inconsistent and unintuitive. Allen achieves the most by attempting the least. He proposes a simple plausible formalism, which makes few ontological or computational commitments. The paper concludes with a high-level discussion of the merits formal logic as a representation for empirical knowledge.

Keywords: knowledge representation, change, time, situation calculus, nonmonotonic logic

This research was supported (in part) by National Institutes of Health Grant No. R01 LM04493 from the National Library of Medicine and National Institutes of Health Grant No. R24 RR01320 from the Division of Research Resources.

Contents

1 Introduction

How can an artificially intelligent computer make it in the real world? It must know what the world is like and use that knowledge to achieve its goals, or in AI terms, it needs a world model on which to base its inference mechanisms. A complete model should capture all important aspects of reality: three-dimensional geometry, topology, objects, properties, naive physics, elementary human psychology, and the epistemological kitchen sink. But that's too much for one paper, so I focus on representing change.

A robot that ignores changes will run into trouble sooner than Mr. Magoo. The first time it tries to cross a momentarily empty street without considering oncoming cars, it will probably (definitely, in Boston) be destroyed. On the other hand, the robot could get nothing done if it assumed that anything could change arbitrarily. It would never try to cross a totally empty street, since a car might appear from nowhere, or the asphalt might disappear. In order to steer between these two extremes, the robot needs to know how and when things change in the real world.

In this paper, I evaluate three recent proposals for representing change and compare them with an earlier one by McCarthy and Hayes. All four use the same strategy, though differing in specifics. They represent the world by sentences of some formal logic and reason about the world within its proof theory. Initially, I adopt this paradigm on faith and examine how well the four models implement it. I discuss McCarthy and Hayes's situation calculus in section 2 and continue with the ideas of Moore, McDermott, and Allen in the next three sections. I argue that all three (despite some claims to the contrary) are simply modified versions of the situation calculus, not radically new ideas. In section 6, I lose my faith and question the formal paradigm itself. According to time-honored tradition, I conclude with a summary.

2 Situation Calculus

McCarthy and Hayes [14] propose a representation for change, called situation calculus, that subsumes similar work by McCarthy [13] and Green

[6]. They utilize the *possible worlds* model of reality developed by Lewis [11,12], Stalnaker [21,22], and other philosophers dating back to Leibniz. Every conceivable instantaneous state of the universe corresponds to a possible world,[1] *situation* in McCarthy's terminology. Successor links connect each situation with all other situations that could follow immediately after it. Every path through the graph of situations forms a possible history of the universe, but only one path traces its actual history.

The situation calculus formalizes the possible worlds theory in first-order predicate calculus. It represents situation dependent propositions and functions as predicates and functions on worlds, called *fluents*. For example, "owns a car" corresponds to the propositional fluent $owns(p, car, s)$ and "president" to the functional fluent *president(s)*. The functional fluent $result(p, a, s)$ denotes the situation that results from agent p performing action a in situation s, so "if you feed a cat it will be quiet" can be formalized as

$$\forall s \forall c \forall agent\, [cat(c) \wedge (t = result(agent, feed(c), s))] \Rightarrow quiet(c, t)$$

where *feed* is a function from objects to actions. Simple actions are primitive irreducible entities, whereas complex actions, called *strategies*, are specified by computer programs in which calls to *result* change the value of a global situation variable. The act of walking five blocks could be written as

```
for i=1 to 5 do s:=result(agent,walk-a-block,s)
```

with s denoting the changing world situation. Unlike the general possible worlds model, situation calculus is deterministic; every action has a unique result. However, nondeterminism could be introduced by generalizing the *result* function to a *possible-result* relation between an agent, an action, a prior situation, and a resulting situation.

McCarthy and Hayes incorporate models of time and knowledge into the situation calculus, without leaving first-order logic. Time is represented by the *time* function from situations to their time of occurrence, a real number, and by the *cohistorical* predicate that holds between pairs of situations belonging to the same world history. Intuitively, the *time* value of a situation

[1] Philosophers disagree on just what possible worlds are. In fact, many deny their existence altogether. For more on this issue, see my paper "What are possible worlds?" [20]

shows when it would have occurred if its history were the actual one. Non-cohistorical situations that have the same *time* represent alternate states of the universe at that instant. *Cohistorical* must be an equivalence relation in any reasonable theory of time, but other axioms may be added as desired. For instance, one could rule out multiple histories by making all situations cohistorical. Other axioms could stipulate continuous, discrete, infinite, finite, or circular models of time. Once *cohistorical* and *time* have been introduced, other useful concepts can be expressed, such as the $\prec$ precedence relation

$$s \prec t \stackrel{\text{def}}{=} cohistorical(s,t) \wedge time(s) < time(t) \qquad (1)$$

which says that situation t lies in the future of s; t occurs in the same history as s, at a later date.

A useful model of the world must represent what agents know, as well as what they do, since the two concepts are interdependent. An ape cannot write this paper because it does not know English, but I can because I do. Conversely, the action of reading this paper affects the readers knowledge. Knowledge can be represented by a first-order analog of Hintikka's [7] logic of knowledge. Hintikka uses a special operator, $know(agent, prop)$ to formalize the proposition "agent knows proposition." The axioms

1. $know(a,p) \Rightarrow p$
2. $know(a,p) \Rightarrow know(a, know(a,p))$
3. $know(a, p \Rightarrow q) \Rightarrow [know(a,p) \Rightarrow know(a,q)]$,

closed under the principle

4. if p is an axiom then $know(a,p)$ is an axiom,

specify the meaning of the *know* operator. Axiom 1 says everything known is true, axiom 2 says everyone knows what he knows, axiom 3 says everyone knows Modus Ponens and principle 4 makes all axioms, including 1–3, common knowledge. Axiom 3 and principle 4 jointly imply that everyone knows all logical implications of his knowledge, including all tautologies, since he knows the axioms and how to apply the rule of inference. For a

fixed agent, these knowledge axioms are equivalent to the modal axioms of $S4$, so adding them to predicate calculus produces a version of modal $S4$ indexed by agents, with a corresponding indexed version of Kripke's [10] possible worlds semantics. Instead of a single accessibility relation, there is one for each agent which holds between two worlds when the first is compatible with everything he knows in the second. The proposition $know(a, p)$ is true in world w iff p is true in all worlds accessible to a from w. This definition, along with the intuitively plausible requirements that the accessibility relation be reflexive and transitive in its world arguments, guarantees the validity of Hintikka's axioms.

Rather than extend situation calculus to include Hintikka's theory of knowledge, McCarthy and Hayes encode its model theory within the existing formalism. The advantage of this approach, discussed at length by Moore [16, pp. 11–20], is the relative simplicity and tractability of standard predicate calculus compared with its modal relatives. They encode the accessibility relation as a reflexive transitive predicate $shrug(a, s, t)$ between an agent and two situations. Two *shrug* related situations are *epistemologically equivalent* to an agent. When asked which of the two he is in, he can only shrug his shoulders. The modal proposition $know(a, p)$ translates to the first-order proposition

$$\forall s\ shrug(a, s_0, s) \Rightarrow p(s) \tag{2}$$

where $p(s)$ is a fluent and s_0 denotes the situation corresponding to the real world. This scheme seems to capture the same intuitions as Hintikka's knowledge axioms, since it translates all of them into theorems of situation calculus. For example, axiom 1 translates to

$$[\forall s\ shrug(a, s_0, s) \Rightarrow p(s)] \Rightarrow p(s_0) \tag{3}$$

which follows from the reflexivity of *shrug* after instantiating s with s_0 and using conditional proof. Hence, McCarthy and Hayes feel justified in replacing Hintikka's modal axioms with their first-order counterparts.

This representation of knowledge seems completely impractical. First, it assumes that agents know all logical implications of their premises, in particular all analytic truths. The standard answer, that completeness is an idealization similar to the frictionless point-mass of mechanics, leaves

me unconvinced. Idealizations should focus attention on key concepts by suppressing irrelevant details. For example, friction distracts a beginner from understanding oscillation or inertia. In our case, completeness is not a useful idealization, since the computational limitations of humans and robots are the concept of interest, not a tedious detail. Similarly, frictionless point-masses are not a useful idealization for modeling torque. Moore [16, p. 8] recognizes this problem, but offers no solution beyond vague references to default reasoning and to the work of Konolige [9]. Konolige's solution is to represent knowledge (or other propositional attitudes, such as belief) in an incomplete deductive system. An agent only knows the implications of its belief in its restricted language. It remains ignorant of their additional logical consequences.

Second, the user must stipulate what situations *shrug* relates, since the axioms merely constrain it to be reflexive and transitive. Moore's work, discussed in the next section, provides a partial solution to this problem by defining *shrug* for the results of an action in terms of its values on prior situations. Yet this still leaves a lot to the user. He must extend the *shrug* relation to each new situation that arises, except for results of actions.

In this section, I have described the situation calculus, a first-order predicate calculus version of the possible worlds model of reality. It represents situation-dependent facts by predicates on situations and change by the *result* function, which maps the performance of an action in a situation to the resulting situation. Theories of time and knowledge can be incorporated into this formalism by defining a few straightforward predicates. In the next three sections, I discuss alternative representations for change, all of which claim to improve upon situation calculus.

3 Moore's Extension

The situation calculus represents actions, time, and knowledge, but fails to tie them together, as McCarthy and Hayes [14, p. 497] realize:

> Rather more interesting [than the familiar axioms of time and knowledge] would be axioms relating 'shrug' to 'cohistorical' and time; unfortunately we have been unable to think of any

intuitively plausible ones. Thus, if two situations are epistemological alternatives (that is, $shrug(p, s_1, s_2)$) then they may or may not have the same time value (since we want to allow that p may not know what the time is), and they may or may not be cohistorical.

Moore [16, sec. 1] stresses the interdependence of knowledge and action in planning. A good planner must reason about knowledge prerequisites for its actions, as well as physical ones, and fill in missing knowledge by taking appropriate actions. For example, a robot cannot open a combination safe unless it *knows* the combination, or make a phone call unless it knows the number. However, McCarthy and Hayes must use the *ad hoc* functions *idea-of-combination* and *idea-of-phone-number* to formalize these rules. Moore proposes a formalism for knowledge and action, in which such examples follow from general principles, without special rules. In section 3.2, I describe a simplified version of his theory that can be expressed directly in situation calculus. Before doing so, I argue that the original gains nothing from its tremendous complexity.

3.1 The Complicated Original

Moore bases his theory of knowledge on Hintikka's modal logic, as do McCarthy and Hayes, but chooses to include both the original object language $\mathcal{H}$ and its model theory in a new first-order object language $\mathcal{M}$, rather than just the model theory. He represents $\mathcal{H}$ variables and constants by $\mathcal{M}$ constants; and $\mathcal{H}$ functions, predicates, and connectives by $\mathcal{M}$ functions, so the $\mathcal{H}$ sentence $know(john, \exists x p(x))$ translates to the $\mathcal{M}$ term, called an *object term*, $know(john, exist(x, p(x)))$ made up of the *know*, *exist*, and p functions and *john* and$x$ constants. Next, he encodes the model-theoretic relation "p is true in world w" as an $\mathcal{M}$ predicate, $t(w, p)$ on situations and object terms. Compound object terms can be reduced to atomic ones by repeated application of appropriate axioms to their main connective, for instance

$$\forall w \forall p \forall q \, t(w, and(p, q)) \Leftrightarrow [t(w, p) \wedge t(w, q)] \tag{4}$$

for conjunction. An atomic object term $p(t_1, \ldots, t_n)$ is true in w iff the denotations of its arguments in w satisfy the corresponding $n + 1$ place

predicate $:p$, according to the axiom

$$\forall w\, t(w, p(t_1 \ldots, t_n)) \equiv :p(w, d(w, t_1) \ldots, d(w, t_n)) \tag{5}$$

where $d(w,t)$ is the denotation of t in w. The d function, in turn, can be reduced to rigid designators, but I skip the details. Finally, t is defined for *know* by the axiom

$$\forall w \forall p \forall s\, t(w, know(p, s)) \equiv \forall u\, k(p, w, u) \Rightarrow t(u, s) \tag{6}$$

where k is a synonym for the *shrug* accessibility relation of section 2.

What advantage does $\mathcal{M}$ have over the simpler knowledge representation of situation calculus? According to Moore,

> This has the advantage of letting us use either the modal language or the possible-world language—whichever is more convenient for a particular purpose—while rigorously defining the connection between the two. [16, p. 30]

I reject this claim. The only connection that Moore defines is between formulas of the form $t(w,p)$ and other, t free, formulas. This connection allows one to interchange propositions about object terms with equivalent object-term-free propositions, but asserts nothing about the relation between an $\mathcal{H}$ sentence and its corresponding object term. To substantiate his claim, Moore must prove some sort of equivalence between $\mathcal{H}$ and $\mathcal{M}$, not between object terms and their definitions.

Although Moore is unaware of his debt to rigor, others have tried unsuccessfully to discharge it. As a first step, it seems reasonable to define the translation of the $\mathcal{H}$ formula h as the $\mathcal{M}$ formula $\forall w\, t(w,m)$ where m is the object term corresponding to h. One might hope that the theorems of $\mathcal{M}$ would be exactly the translations of the theorems of $\mathcal{H}$. Unfortunately, this proves false, since $\mathcal{M}$ contains theorems, such as $\forall w\, w = w$ that are not the translations of any $\mathcal{H}$ formula, let alone of theorems. The conjecture

Conjecture 1 *h is a theorem of $\mathcal{H}$ iff its translation is a theorem of $\mathcal{M}$*

avoids this difficulty and seems plausible. However, I see no way of proving it, nor have I found one in the philosophical literature. In fact, Lewis

[12] proposes a translation of modal logic into first-order predicate calculus that closely resembles $\mathcal{M}$, but makes no equivalence claims. More recently, Forbes [5, chap. 4] discusses the difficulties involved in proving conjecture 1. In any case, the burden of proof lies with Moore.

Theoretical difficulties aside, I think $\mathcal{M}$ is a bad idea, because it confuses data structures with algorithms. A theory of knowledge, like any database, should consist of two parts: a representation for knowledge and algorithms for translating input from external to internal form and back. Moore uses the knowledge representation to do both jobs, thus making his system needlessly complex. Each knowledge query must be translated into situation terms, analyzed, and the results translated back into modal language, as demonstrated by the theorems proved by Moore [16, pp. 39–40, 61–64]. It is much simpler to keep the first-order model free of any reference to the modal language and leave translation to an interface program. The new formalism cannot express the (as yet unproved) equivalence between $\mathcal{H}$ formulas and their object terms, but that job belongs to the user interface. Put another way, the equivalence should be proved and used in the meta-language[2] that describes the knowledge representation, not in the object language.

In this section, I criticized Moore's first-order translation of Hintikka's logic of knowledge $\mathcal{H}$ for being overly complex and suggested simplifications. The simplified formalism is equivalent to the situation calculus theory translation of $\mathcal{H}$ described in section 2. In the next section, I recast Moore's theory of knowledge and action in situation calculus, thus demonstrating that the simplified theory retains sufficient expressive power for his purposes. His only justification for the more complicated one, that modal propositions be interchangeable with their translations, confuses object language and meta-language.

3.2 A Simpler Alternative

Moore's primary goal is to formalize the interdependence of knowledge and action, which eluded McCarthy and Hayes. He reinterprets the action argument of the *result* function as an *event*. This distinction has no practical

[2] Every paper needs a meta.

effect on the theory, since the only events considered consist of actions by agents, represented by the $do(agent, action)$ function. However, it reappears in sections 4 and 5, so keep it in mind. Just as in situation calculus, complex actions can be specified by sequences of actions, conditional actions, and iteration over actions. Each has an appropriate event associated with it by the *do* function.

How does action depend on knowledge? Moore claims that knowing how to perform an action consists of having a descriptor for that action and knowing what object it denotes. Moreover, most routine actions can be described as applications of known procedures to appropriate arguments. In these cases, knowing how to perform the action reduces to knowing its arguments. For example, a combination safe can be opened by the general routine *dial(combination(safe),safe)*, so knowing how to open it reduces to knowing the denotations of *safe* and *combination(safe)*, that is knowing which safe to open with what combination. Knowing what an expression denotes in world w is expressed in situation calculus as

$$\exists x \forall u \; k(agent, w, u) \Rightarrow x = exp \tag{7}$$

where *exp* contains fluents with world variable u. The intuitive meaning of (7) is that *exp* denotes the same object in every world compatible with what *agent* knows in w. In particular, a burglar could open a safe if the proposition

$$\exists x \forall u \; k(burglar, real\text{-}world, u) \Rightarrow x = combination(safe(u), u) \tag{8}$$

were true. An agent can also know how to perform an action without initially knowing what object it denotes, if he knows a multiple-step action whose early steps fulfill the knowledge prerequisites of its later ones. Moore claims that these are the only two ways.

An agent's actions affect its knowledge, just as its knowledge affects its actions. Moore distinguishes between *informative* actions that tell the agent something about the world and *noninformative* actions that do not. Reading a thermometer is informative, whereas dancing is not. (Of course these are abstractions, since every real-world action yields some information.) Let us call the situations compatible with an agent's prior knowledge *antecedents* and those compatible with his post-action knowledge *consequences*. After performing a noninformative action, an agent knows that it

occurred, but learns nothing else, so each consequence is the result of some antecedent. Equivalently, the consequences are the image under *result* of the antecedents. This is expressed formally as

$$\forall u \forall v \; result(e,u,v) \Rightarrow [\forall w \; k(a,v,w) \Leftrightarrow \exists x \; k(a,u,x) \wedge result(e,x,w)] \quad (9)$$

a state w is compatible with the result of event e in state u iff w is the result of e in some state x compatible with u. On the other hand, after performing a p informative action, an agent knows whether or not p is true, so the consequences are those results of antecedents in which p takes on the correct truth value. This is formalized by adding a third conjunct $p(v) \equiv p(w)$ to (9). As an example, suppose w_3 and w_4 are the results of e on w_1 and w_2 and that $p(w_3)$ is true and $p(w_4)$ false. If e is noninformative and w_1 and w_2 are John's antecedents then w_3 and w_4 are his consequences. If e informs John that p is true then w_3 is his only consequence.

3.3 Summary

Moore hypothesizes relations between knowledge and action and formalizes them in situation calculus. An agent knows how to perform an action if it has a descriptor for the action and knows what action the descriptor denotes. Performing an action maps an agent's state of knowledge to one compatible with the results and the agent's prior knowledge. Moore improves on McCarthy and Hayes's work by binding together two formerly independent concepts into a single integrated theory.

4 McDermott's Confusion

McDermott [15] proposes a temporal logic for representing causality, continuous change and problem solving. He prefaces the theory with a bold claim:

> Actually, of course, no one has ever dealt with time correctly in an AI program.

According to him, the underlying logic of existing programs, situation calculus, cannot adequately model time-varying facts or hypothetical future

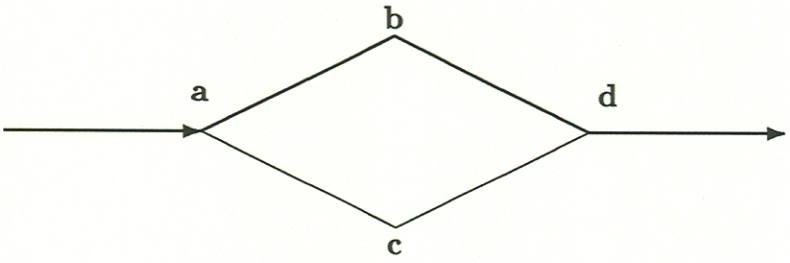

Figure 1: Convexity Counterexample

events. In this section, I argue that his theory is logically confused, philosophically dubious and overly complex. Nor does his case against situation calculus seem compelling, since time-varying facts can be represented by fluents and hypothetical futures by situations.

4.1 Time, Facts, and Events

McDermott adopts the situation calculus model of states partially ordered by the $\prec$ relation and mapped onto the real line by the *time* function. (In fact, he adopts the nonstrict version of $\prec$, which he renames $\leq$, and *time* as his primitives and defines *cohistorical* in terms of them.) Whenever one state strictly precedes another, some third state lies in between. Next, he defines *chronicles,* totally ordered sets of states that cover all times, symbolically

$$\forall t \exists s \; s \in ch \wedge time(s) = t \tag{10}$$

where t is a time, s a state, and ch a chronicle. Each chronicle represents a complete possible history of the universe. An ordered set is *convex* if whenever it contains two points, it contains all points in between. Figure 1 provides a counterexample to McDermott's claim that chronicles are convex by their definition. State c is between a and d, but outside the abd chronicle.

Fortunately, convexity follows from the restriction that states never branch into the past

$$\forall s \forall t \forall u (s \leq u \wedge t \leq u) \Rightarrow (s \leq t \vee t \leq s) \tag{11}$$

only the future, because there exists only one, perhaps unknown, past, but many possible futures. Suppose state s belongs to chronicle ch and t strictly precedes s. There exists $t' \in ch$ such that $time(t') = time(t)$ by (10). State

t' precedes s because $time(t') = time(t) < time(s)$ and t precedes s by hypothesis, so either t' precedes t or t precedes t' by (11). Either way, the precedence cannot be strict because $time(t) = time(t')$, so they must be equivalent.

Axiom (11) reflects a genuine asymmetry of time: actions can change the future, but not the past. It allows an agent to focus on planning for the future without wasting time constructing an overly general model of the past. However, I think this distinction should be enforced semanticly from outside the formalism, not syntactically from within. In some cases, an agent might want to consider branching pasts—perhaps in reasoning backwards from dinosaur fossils to situations in which they lived—or non-branching futures. This decision should be made by the agent based on his current needs, not fixed once and for all by axiom (11). I disapprove of branching futures too, but postpone discussion to section 5.

Having specified a formalism for time, McDermott turns to facts, events, and causality. A *fact* is a set of states and an *event* a set of intervals: convex sets of states. A fact is true in each of its states, while an event occurs exactly once over each of its intervals with no time left over on either side. The intervals belonging to an event are its instances in all possible chronicles. Facts obey a generalized continuity principle: for every fact f and situation s in chronicle ch there are *intervals* preceding s and following s which are contained in f or its complement. In other words, a fact cannot change truth-value infinitely often in the neighborhood of any state.

These definitions of facts and events are philosophically problematic, since they reduce intensional concepts to extensional ones. Surely, a fact or event is not identical with its times of occurrence. If that were true, contemporaneous facts and events would be identical. For example, every time Alexander makes a local phone call he is charged one message unit, but making local phone calls and spending message units are different events. Alternately, suppose a bus always leaves from Boston at the same time a train leaves from Davenport, and arrives in New York City at the same time the train pulls into Grand Rapids. Two distinct events still occur. Philosophers disagree on whether distinct events can really have the same instances in all possible situations, but from a *pragmatic* point of view the answer is clear. A real program can only represent a countable subset, probably only a finite number, of the uncountably many possible states.

For this reason, it should never assume that two events that always occur together in its model are truly cotemporaneous. The bus and the train may very well travel together in all represented states, but these trips must never be reduced to the same event. McDermott's extensional model of events and facts encourages such mistakes.

McDermott claims that situation calculus cannot represent the duration and intermediate results of events, since its *result* function maps initial situations directly to the final results of actions. I find these charges unconvincing, since situation calculus defines the $\prec$ relation and *time* function. One can represent the duration of an event by the difference between the times of its final and initial situations and intermediate events by situations that follow the initial state and precede the final one, just as McDermott does. A second claim, that events that are not fact changes cannot be expressed at all, deserves more thought. From a formal point of view, this is false, since his model is a restricted version of situation calculus: it contains the same *time* and $\prec$ along with additional axioms, such as (11). Anything that can be expressed in the restricted system can be expressed in the more general one as well. However, from a pragmatic perspective, I agree that a useful world model must provide the user with a richer model of events than does the situation calculus. Although McDermott thinks his work is a revolutionary alternative to situation calculus—as he makes clear in his introduction—it is really an evolutionary extension, which adds structure to the existing framework.

4.2 Causality

Causality between events is expressed by the $ecause(pre, c, e, d)$ predicate, meaning that effect e occurs within delay d after cause c if condition pre holds during that period. Causation is a primitive irreducible aspect of reality, so there are no special inference rules for deriving instances of *ecause* from other facts. If any instance of an event has a cause then all do. In my opinion, this latter principle confuses the formalism with reasoning methods, as does axiom (11). Whether or not a user assumes that all events have causes should be up to him, not the syntax of his world model. In fact, it might be false for some useful and intuitive theories of causality. For example, consider two coins, the first loaded and the second fair, each

of which comes down heads 100 times consecutively. According to our intuitions, the first event was caused by the design of the coin, whereas the second was pure chance.

At first glance, it seems natural to represent causality between an event and a fact analogously to causality between two events. If an event causes a fact then the fact remains true for some period after the event ends. However, McDermott finds this definition unacceptable because it says nothing about the fact's duration, or in his words

> We could do this, but it would be useless. In this sense, shooting a bullet past someone would be a way of achieving that it was near him. [15, p. 119]

I reject this view. Shooting a bullet past McDermott certainly does cause it to be near him, albeit for a short while. An event causes a fact by leading to a state in which the fact is true. How long that state persists is another matter, depending partially, but not solely, on the causing event.

After rejecting causality between events and facts, McDermott develops a theory in which an event causes the *persistence* of facts for a given period. He thinks it perfectly reasonable to reify persistences, since "the senses actually tell you about persistences." If people only sensed facts, they would suffer from the frame problem, since what is true now might well be false in an instant. They avoid that famous bugaboo of AI by sensing persistences. I find this unconvincing: it is pure unsubstantiated speculation without explanatory power. If sensing persistences just means sensing that facts persist, McDermott's theory merely postulates a solution to the frame problem and names it persistence, without explaining anything new. If it means that people sense entities called persistences, he owes us evidence of their existence and an account of how they solve the frame problem. Next, I evaluate the role of persistences in causality from an artificial intelligence perspective.

The $persists(s, p, r)$ predicate means that fact p is true in state s and if p is false for any state t within r time of s then a $cease(p)$ event occurs between s and t. For instance, $persists(now, on(light), 1\ hour)$ expresses the proposition that the light will be on for an hour unless something extinguishes it. The $cease(p)$ event occurs iff it is *inconsistent* that it not

occur. With this nonmonotonic definition of persistence in hand, McDermott defines the *pcause* predicate, which states that an event causes the persistence of a fact for a certain period.

McDermott's use of nonmonotonicity raises theoretical and practical difficulties. It contradicts his declared motive for proposing a logic-based model in the first place.

> So why do I plan to spend any time at all on logic?... We want to be assured that our special-purpose modules are not prone to absurd interactions... One way to guarantee this is to be sure that the modules' actions are sound with respect to an underlying logic. [15, p. 103]

That reasoning makes sense for a well-behaved logic like predicate calculus, where a reasoner can apply rules of inference to axioms and premises to derive new theorems. The soundness theorem prevents him from ever coming up with incorrect results. However, nonmonotonic logics have no deduction rules or other local proof methods. It is impossible in general to determine whether a formula follows from premises without generating an entire, necessarily infinite, fixed-point of the theory. Worse yet, even the most benign premises may have several disparate fixed-points or none at all. A guarantee that all deductions are correct means little in a logic where none can be made. As McDermott admits,

> I try to appeal to nonmonotonic deductions as seldom as possible. This is because the logics they are based on are still rather unsatisfactory. For one thing, even some of the simple deductions in this paper may not be valid in any existing nonmonotonic system. [15, p. 122]

McDermott believes nonmonotonic logic can be made workable by logicians, if AI practitioners explain what types of inference they need. He treats its shortcomings "not as problems with this paper, but as problems with nonmonotonic logic." Unfortunately, his optimistic view runs counter to Moore's [17, pp. 77–79] argument that no sound logic can include default reasoning because it is an inherently unsound rule of inference. Any attempt to include defaults in a logic must fail because default rules sometimes derive incorrect results from correct premises, but logics never do.

McDermott defeats his original goal of soundness by introducing nonmonotonicity.

In my opinion, the whole attempt to introduce nonmonotonicity stems from the confusion between a model and its use, which I have mentioned before. A reasoner may revise his theory of the world nonmonotonically by deleting old beliefs based on new information. This metal-level operation need not be sound in a logical sense—in fact, unsound revisions probably yield the best results—but the resulting theory must be logically sound, if its theorems are to be useful. Hence, any attempt to express the unsound meta-level objectives in the sound object language are misguided and futile.

McDermott repeats[3] Moore's theory of action: the *do(agent,action)* function maps a performance of *action* by *agent* to an event. He composes compound actions, called *plans*, from sequences of simple actions, but eschews more powerful constructs, such as iteration and local variables. Unlike Moore and McCarthy, he realizes that the complexity of program-like actions makes it impractical to represent them as deduction rules of predicate calculus. It can be done, but mechanical, or human, reasoners have little chance of understanding the results. Perhaps a different logic, such as Pratt's [19] dynamic logic, would represent action better than predicate calculus, but I cannot pursue that topic for lack of time and space.

Next, McDermott defines useful types of action, such as allowing, preventing, forgoing and bringing about. I only describe preventing since the rest are similar. By definition, action a prevents event e if e depends negatively on the event *do(agent,a)*. Event e_2 depends negatively on e_1 if e_2 occurs in some chronicle, does not occur in every chronicle and never occurs in the same chronicle as e_1. Although this account may be theoretically adequate, its implementation confuses correlation with causation. As discussed earlier, the agent can only represent a tiny fraction of the possible situations. It should not assume that event e_2 depends negatively on e_1 just because they happen never to occur jointly in its data base. Furthermore, this definition ignores time of occurrence. By ringing today, Brook's phone prevents the cord from being cut yesterday, since the cutting and the ringing could not share a chronicle. Yet cutting the cord cannot prevent the

[3]Moore's 1985 paper [16] reviews ideas from his 1980 doctoral thesis, preceding McDermott by several years.

phone ringing, because it might have rung yesterday. In trying to represent prevention, a causal relation, as correlation, McDermott has ignored his own advice, "there is no way to infer causality merely from correlation [15, p. 117]."

In a personal communication, Allen charges that real-world actions never satisfy McDermott's prevention and achievement criteria. I cannot really prevent the phone from ringing by destroying it, since some brilliant ex-employee of Western Electric might reconstruct it. Nor, to use McDermott's example, can Dudley prevent the train from crushing Nell by untying her. She might jump back on the tracks and commit suicide. In general, a problem-solver cannot *prove* that any plan achieves or prevents anything without restricting its world model unduly. It must choose the plan most likely to succeed and ignore improbable obstacles. McDermott hopes that "a dose of non-monotonicity" will enable his system to prove things based on default assumptions. He concedes that it cannot represent probabilistic information necessary for choosing plans likely to succeed and ignoring rare events, but claims that this problem transcends his work to encompass all AI research. I return to this issue in section 6.

4.3 Summary

McDermott rejects situation calculus and proposes a nonmonotonic logic for time, change, and action. His model of time resembles situation calculus, except that time cannot branch into the past, while his model of action resembles Moore's, except that program-like actions are banned. He equates facts and events with the time of their truth and occurrence, leading to philosophical and practical problems. He reifies persistences of facts and relates them to events by a nonmonotonic axiom. I find this approach unpromising because of the inadequacies of nonmonotonic logics. Finally, his definitions of prevention and achievement are unintuitive and impractical. In the next section, I describe a better solution to the problems McDermott addresses.

5 Allen's Simplification

Allen [1] finds situation calculus inadequate for representing:

- nonactivity: standing on the corner for an hour,
- nondecomposable actions: spending the day hiding from someone, and
- simultaneous interacting action: walking to the store while juggling.

As I showed on page 13, this is a claim about expressive power, not logical adequacy. In principle, situation calculus can represent the same things as Allen's formalism, since they are both extensions of predicate calculus. Theoretical equivalence, though, does not imply practical equivalence. A formalism that captures explicitly the important aspects of a domain facilitates reasoning, whereas a formalism that leaves them implicit causes confusion. For this reason, Allen designs his model of time and action to represent explicitly the domain's key concepts: time, event, process, and action.

5.1 Properties, Events, and Processes

Allen uses an interval-based temporal logic, described in [2], general enough to model discrete time, continuous time, and combinations of the two. By and large, the precise theory does not affect our topic, so I will just equate his intervals with the standard concept of a time interval. However, there is one exception to this rule: Allen rules out time points and, consequently, identifies open, half open, and closed intervals. He finds points unnecessary, since they can be viewed as tiny intervals. I disagree. Time points are useful and important concepts. When does a pendulum reach the end of its swing, if not at a particular instant? Of course, one man's points are another microscope's intervals, but points are useful at any fixed level of abstraction, regardless of their ultimate ontological status. Allen [2] himself admits as much when he reinvents the wheel and constructs points out of intervals.

Nor do I agree with Allen's argument that introducing points leads to inconsistency or truth gaps:

> For example, consider the time of running a race, R, and the time following after the race, AR. Let P be the proposition representing the fact that the race is on; P is true over R and $\sim$ P is true over AR. We want AR and R to meet in some sense. Whether both ends of the intervals are open or closed, AR and R must either share a time point or allow time between them. Thus we have a choice between inconsistency or truth gaps, i.e., either there is a time when both P and $\sim$ P are true, or there is a time when neither P nor $\sim$ P is true. [1, p. 128]

All this argument shows is that intervals divide into four similar subclasses: closed, open, left open and right open. These distinctions can often be ignored, but sometimes they must be made. In the race case, we may put the boundary point b between R and AR into either R, AR both or neither. The predicate P(b) is true iff b is in R. Allen's purported truth gap and inconsistency result from a confusion between $\neg$P(t), which equals $t \notin$ R, and $t \in AR$. These are the same only if b belongs to exactly one of A and AR.

Adopting the ontology of Mourelatos [18], he divides time-dependent aspects of the world into *properties* and *occurrences*. Properties represent persistent facts about objects, such as "the house was warm all Monday." A property holds over an interval, denoted by *holds(p,int)*, iff it holds over every subinterval. Thus, the house was warm all Monday iff it was warm over ever subinterval of Monday. Although I accept this definition in principle, I think it fails in Allen's interval-based time, providing another argument for points. Some intuitively reasonable properties hold over an open interval, but not its closure. This situation contradicts Allen's condition, so he must impoverish the world model by ruling out such properties. For example, consider a ball that is thrown straight up at time t_0 and reaches its apex at time t_m. It has the property of rising on every subinterval of $[t_0, t_m)$, but not on the closed interval. Allen must either say it does, or deny that rising is a property.

Occurrences model dynamic aspects of the world. They divide into two subclasses: *events* describe discrete activities with definite endpoints, whereas *processes* describe continuous activities without definite outcomes.[4]

[4]McDermott's events lump these events and processes together. In order to represent

As a general heuristic, one can count events, not processes. "He bought a coat" refers to an event and "the ball is falling," to a process. It makes sense to ask how many times he bought a coat, but not how many times the ball is falling. In contrast, "the ball fell to the ground" is an event because it has a clear outcome, so its instances can be enumerated. In Allen's terminology events *occur* and processes are *occurring*. In light of the counting principle, if an event occurs over interval *int* then it does not occur over any subinterval of *int*. Allen does not know exactly what a process's occurring over an interval implies about its subintervals. He stipulates that it must be occurring over *some* subinterval, for lack of a stronger condition. Such a condition would formalize the intuition that the process is occurring on most subintervals, or on a large enough percentage to guarantee continuity. Like Allen, I see no general way to formalize this.

Allen formalizes causation between events by an *ecause* predicate analogous to McDermott's. If event c occurs over t_c and $ecause(c, t_c, e, t_e)$ is true then event e occurs over t_e. Backward causations cannot occur, that is t_c must be no later that t_e. Event causation is transitive, anti-reflexive and anti-symmetric, corresponding to our intuitions:

1. if a causes b and b causes c then a causes c,
2. nothing causes itself, and
3. two events cannot cause each other.

It is a primitive concept irreducible to others, so new causal relations can only be deduced from existing ones, nothing else. I think this decision, also made by McDermott, is a wise one. Causal relations should be deduced by examining a world model from outside, not proved from within, since they are never deductively valid, only inductively reasonable.

5.2 Actions and Plans

Actions are a subset of occurrences, as in situation calculus, not a separate class of entities, as postulated by Moore and McDermott. The function

Allen-processes, McDermott allows all events to occur over an interval and its subintervals. This leads to confusion and mistakes in the case of Allen-events that *cannot* occur over their subintervals.

acause(*agent*, *occ*) returns the action of *agent* causing *occ*. This action is an event, called a *performance*, if *occ* is an event and a process, called an *activity*, if *occ* is a process. If an agent acauses an event or process over a time interval, that event or process occurs or is occurring respectively over the interval. Many actions consist of agents acausing occurrences. In fact, Allen conjectures that all actions can be described in this way fairly naturally.

Unlike McCarthy, Moore, and McDermott, Allen adopts a single time line, rather than branching time. He performs hypothetical reasoning from outside the formalism by creating and analyzing hypothetical models, rather than encoding the mechanism within the formalism. Agents plan by manipulating three partial world descriptions: the expected world, desired world, and planned world. The *expected* world contains what the agent believes will happen if certain known events occur, he does nothing, and everyone else does as little as possible. In Allen's words,

> The expected world obeys a generalized law of inertia. Things in the process of changing continue to change unless prevented, and everything else remains the same unless disturbed. [1, p. 144]

The *desired* world consists of the properties, events, and processes that the agent desires. In order to achieve its goals, the agent forms a *plan*, a set of decisions to perform actions not in the expected world and refrain from actions in the expected world. Applying a plan to the expected world produces a *planned* world, a simulation of the plan's effects on the expected world. Loosely speaking, an optimal plan for an agent minimizes the differences between the corresponding planned world and his desired world. Allen suggests a generalized GPS model of difference reduction as a planning algorithm, but defers discussion to another paper [3]. My topic is modeling not planning, so I do the same.

5.3 Summary

Allen prefers an interval-based linear model of time to possible worlds. He adopts an ontology of properties, events, and processes and represents them explicitly in his formalism. This improves on the implicit representation of

McCarthy and Hayes and Moore's partial one, in which all three categories are lumped together. Actions are special cases of occurrences, instances of agents causing occurrences, not a separate type of entity. The linear time line simplifies the model, thus facilitating inference, without harming hypothetical reasoning. That is done from outside the formalism, by constructing hypothetical worlds.

6 Questioning the Paradigm

So far, I have compared the expressive power of four representations for time-varying behavior as logical theories. Now, I turn to their common strategy, expressed by McCarthy and Hayes

> we want a computer program that decides what to do by inferring in a formal language that a certain strategy will achieve its assigned goal. [14, p. 463]

They model the world as a logical theory in order to reduce reasoning about the world to formal deduction, a mechanizable chore. Action reduces to theorems about the *result* function, knowledge to *shrug* relations, causation to *ecause, pcause,* or *acause*, and so on. In each case, the soundness of the underlying logic guarantees that all deductions are valid.

I find this approach unpromising. Syntactic theorem proving methods, such as resolution or natural deduction, suffer from combinatorial explosion in toy domains, let alone realistic world models. Semantic information must be used to guide the process, as Bledsoe [4] demonstrates for mathematics. I came to the same conclusion after taking courses in mathematical logic. The best way to prove a theorem is by examining its interpretation in an appropriate model, such as the integers, rationals, or reals. An informal proof in that interpretation often transforms directly into a formal proof, or at least suggests a line of attack.

Well then, why not use semantic information to guide deduction about world models? Therein lies the tale. Most, if not all, real world reasoning employs unsound inference schemes, such as induction, intuition, and defaults. Its hard to see how these transfer to a sound logic. For example, the default rule "dogs are friendly" allows me to walk home without fearing

attack. Yet, my rule cannot be part of a sound inference system since it has incorrect implications—witness the scar on my left calf. Worse yet, human theories of the world contain inconsistencies. Our informal inference techniques accommodate minor flaws that would destroy a formal theory. I never draw arbitrary conclusions from the contradictory propositions "dogs are friendly" and "there exists a vicious dog" contained in my world model, but a standard theorem prover does. In section 4.2, I argue that nonmonotonic reasoning cannot represent unsound default rules, such as "most dogs are friendly," because it is sound. Hence, it cannot avoid contradictions by replacing general rules with default rules.

The dichotomy between unsound informal inference and sound formal inference explains most of my previous criticisms of the four formalisms. People reason about actions without knowing their exact effect on the entire universe, but those inferences cannot be captured in situation calculus. Some supposedly irrelevant factor might cause them to fail, whereas infallible inference rules alone can be expressed by the *result* function. Knowledge is deductively closed in the formalisms of McCarthy and Moore, but not in people, computers, or any finite beings. In all four systems, agents construct plans that provably achieve their goals, while humans make do with plans that succeed unless something unusual happens.

Am I rejecting deduction outright? Not at all, just putting it in perspective. As Israel [8] explains, formal logic is just one part of human reasoning. Deduction is great at proving tautologies and other *analytic* truths, but poor at deriving *synthetic*, fact-dependent, truths. Proponents of formal theories should use it accordingly, as a tool for exploring the logical connections between facts and assumptions, not an all encompassing model of reasoning. Such a model must include the nondeductive inference methods, such as induction and default reasoning, upon which most real world reasoning depends. They cannot be encoded as inference rules of a formal logic.

7 Conclusions

In this paper, I discuss four formalisms for representing events, actions, and other time-varying information. The first, proposed by McCarthy and

Hayes, formalizes an ontology of possible worlds, called situations, in first-order predicate calculus. Predicates and functions on situations represent time-dependent facts and actions. The *result* function maps every situation s and action a to the unique situation resulting from a's happening in s. McCarthy and Hayes include a theory of time and a first-order translation of Hintikka's logic of knowledge in their formalism. Moore constructs an unnecessarily complicated variant of this theory of knowledge and uses it to describe the relations between knowledge and action. However, his results can be formulated in the simpler situation calculus formalism just as well. McDermott adds events and facts to the situation calculus formalism, increasing its expressive power. Unfortunately, he represents them by their extensions, the times at which they hold, creating philosophical and practical problems. Worse yet, he defines nonmonotonic persistences of facts, thereby confusing the formalism and destroying its proof theory. Allen takes the opposite approach. He replaces the ontology of situations with a linear time order and transforms hypothetical reasoning from an object-level deduction mechanism to a meta-level, possibly nondeductive, form of inference. He improves on McDermott's representation of events by subdividing them into properties, events, and processes and restoring the distinction between an occurrence and the time at which it occurs. He also simplifies Moore's and McDermott's representation of actions by reducing them to occurrences.

Of the four theories, I find Allen's clearest and most expressive. However, it does not include a representation for knowledge and belief, reportedly for lack of space. In any case, formal logic alone cannot represent human reasoning, since the former is sound, whereas the latter need not be. I predict that attempts to ignore this distinction and reduce all reasoning to deduction will fail.

References

[1] James F. Allen.
Towards a general theory of action and time.
Artificial Intelligence, 23:123–153, 1984.

[2] James F. Allen and Patrick J. Hayes.

A common-sense theory of time.
In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 528–531, IJCAI, Los Angeles, California, August 1985.

[3] James F. Allen and Johannes A. Koomen.
Planning using a temporal world model.
In *Proceedings of the Eigth International Joint Conference on Artificial Intelligence*, pages 741–747, IJCAI, Karlsruhe, West Germany, August 1983.

[4] W. W. Bledsoe.
Non-resolution theorem proving.
Artificial Intelligence, 9(1):1–35, 1977.

[5] Graeme Forbes.
The Metaphysics of Modality.
Clarendon Library of Logic and Philosophy, Clarendon Pres, Oxford, 1985.

[6] C. Cordell Green.
Theorem-proving by resolution as a basis for question-answering systems.
In *Machine Intelligence 4*, pages 183–205, Edinburgh University Press, Edinburgh, 1969.

[7] J. Hintikka.
Knowledge and Belief.
Cornell University Press, Ithaca, New York, 1962.

[8] David J. Israel.
The role of logic in knowledge representation.
Computer, 16(10):37–41, 1983.

[9] Kurt Konolige.
Belief and Incompleteness.
Technical Note 319, SRI International, January 1984.
appears in *Formal Theories of the Common-Sense World*, edited by Jerry Hobbs.

[10] Saul Kripke.

Semantical considerations on modal logic.
In L. Linsky, editor, *Reference and Modality*, pages 63–72, Oxford University Press, London, England, 1971.

[11] David Lewis.
Counterfactuals.
Blackwell, Oxford, 1973.

[12] David Lewis.
Counterpart theory and quantified modal logic.
In *Philosophical Papers*, pages 26–46, Oxford University Press, New York Oxford, 1983.

[13] John McCarthy.
Situations, actions and causal laws.
Memo 2, Stanford Artificial Intelligence Project, 1963.

[14] John McCarthy and Patrick Hayes.
Some philosophical problems from the standpoint of artificial intelligence.
In *Machine Intelligence 4*, pages 463–502, American Elsevier Publishing Compan, Inc., New York, 1969.

[15] Drew McDermott.
A temporal logic for reasoning about processes and plans.
Cognitive Science, 6:101–155, 1982.

[16] Robert C. Moore.
A formal theory of knowledge and action.
Technical Report CSLI-85-31, Center for the Study of Language and Information, Stanford University, September 1985.

[17] Robert C. Moore.
Semantical considerations on nonmonotonic logic.
Artificial Intelligence, 25:75–94, 1985.

[18] A.P.D. Mourelatos.
Events, processes and states.
Linguistics and Philosophy, 2:415–434, 1978.

[19] V. R. Pratt.
The role of logic in computer engineering.

1977.
Available in my files.

[20] Elisha Sacks.
What are possible worlds?
Please ask for a copy. I'll be grateful.

[21] Robert Stalnaker.
Possible worlds.
Nous, 10:65–75, 1976.

[22] Robert Stalnaker.
A theory of conditionals.
In W. L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*, pages 41–55, Basil Blackwell, 1968.