Intensions in Communication
MIT Press 1990
Cohen, Morgan + Pollack (eds)

# Chapter 3

# Persistence, Intention, and Commitment

*Philip R. Cohen and Hector J. Levesque*

## 1 Introduction

This paper is concerned with specifying the "rational balance"[1] needed among the beliefs, goals, plans, intentions, commitments, and actions of autonomous agents. For example, it would be reasonable to specify that agents should act to achieve their intentions and that agents should adopt only intentions that they believe to be achievable. Another constraint might be that once an agent has an intention, it believes it will do the intended act. Furthermore, we might wish that agents keep (or commit to) their intentions over time and not drop them precipitously. However, if an agent's beliefs change, it may well need to alter its intentions. Intention revision may also be called for when an agent tries and fails to fulfill an intention, or even when it succeeds. Thus, it is not enough to characterize what it means for an agent to have an intention; one also needs to describe how that intention affect the agent's beliefs, commitments to future actions, and ability to adopt still other intentions during plan formation.

Because autonomous agents will have to exist in *our* world, making commitments to us and obeying our orders, a good place to begin a normative study of rational balance is to examine various commonsense relationships among people's beliefs, intentions and commitments that seem to justify our attribution of the term "rational." However, rather than

This research was made possible in part by a gift from the System Development Foundation, in part by support from the Natural Sciences and Engineering Research Council of Canada, and in part by support from the Defense Advanced Research Projects Agency under contract N00039-84-K-0078 with the Naval Electronic Systems Command. The views and conclusions contained in this document are those of the authors and should not be interpreted as representative of the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States government (or the Canadian one, for that matter).

James Allen, Michael Bratman, Joe Halpern, David Israel, Ray Perrault, and Martha Pollack provided many valuable suggestions. Discussions with Doug Appelt, Jim des Rivières, Michael Georgeff, Kurt Konolige, Amy Lansky, Joe Nunes, Calvin Ostrum, Fernando Pereira, Stan Rosenschein, and Moshe Vardi have also been quite helpful, Thanks to you all.

just characterizing agents in isolation, we propose a logic suitable for describing and reasoning about these mental states in a world in which agents will have to interact with others. Not only will a theorist have to reason about the kinds of interactions agents can have, agents may themselves need to reason about the beliefs, intentions, and commitments of other agents. The need for agents to reason about others is particularly acute in circumstances requiring communication. In this vein, the formalism serves as a foundation for a theory of speech acts (Cohen and Levesque 1985 and chapter 12 of this volume) and applies more generally to situations of rational interaction in which communication may take place in a formal language.

In its emphasis on formally specifying constraints on the design of autonomous agents, this paper is intended to contribute to artificial intelligence research. To the extent that our analysis captures the ordinary concept of intention, this paper may contribute to the philosophy of mind. We discuss both areas below.

### *1.1 Artificial Intelligence Research on Planning Systems*

Artificial intelligence research has concentrated on algorithms for finding plans to achieve given goals, on monitoring plan execution (Fikes and Nilsson 1971), and on replanning. Recently, planning in dynamic, multi-agent domains has become a topic of interest, especially the planning of communication acts needed for one agent to affect the mental state and behavior of another (Allen 1979; Allen and Perrault 1980; Appelt 1981; Cohen and Levesque 1980, 1985; Cohen and Perrault 1979; Georgeff 1983; Georgeff and Lansky, in preparation; Konolige and Nilsson 1980; Rosenschein 1986; Rosenschein and Genesereth 1984). Typically, this research has ignored the issues of rational balance—of precisely how an agent's beliefs, goals, and intentions should be related to its actions.[2] In such systems, the theory of intentional action embodied by the agent is expressed only as code, with the relationships among the agent's beliefs, goals, plans, and actions left implicit in the agent's architecture. If asked, the designer of a planning system may say that the notion of intention is defined operationally: a planning system's intentions are no more than the contents of its plans. As such, intentions are representations of possible actions the system may take to achieve its goal(s). This much is reasonable; there surely *is* a strong relationship between plans and intentions (Pollack, this volume). However, what constitutes a plan for most planning systems is itself often a murky topic.[3] Thus, saying that the system's intentions are the contents of its plans lacks needed precision. Moreover, operational definitions are usually quite difficult to reason with and about. If the program changes, then so may the definitions, in which case there would not be a fixed set of specifications that the program implements. This paper can be seen as providing both a logic in which to write specifications for autonomous agents and an initial theory cast in that logic.



### *1.2 Philosophical Theories of Intention*

Philosophers have long been concerned with the concept of intention, often trying to reduce it to some combination of belief and desire. We will explore their territory here, but we cannot possibly do justice to the immense body of literature on the subject. Our strategy is to make connection with some of the more recent work and hope our efforts are not yet another failed attempt, amply documented in The Big Book of Classical Mistakes.

Philosophers have drawn a distinction between future-directed intentions and present-directed ones (Bratman 1984, 1987; Searle 1983). The former guide agents' planning and constrain their adoption of other intentions (Bratman 1987), whereas the latter function *causally* in producing behavior (Searle 1983). For example, one's future-directed intentions may include cooking dinner tomorrow, and one's present-directed intentions may include moving an arm now. Most philosophical analyses have examined the relationship between an agent's doing something intentionally and that agent's having a present-directed intention. Bratman (1984) has argued that intending to do something (or having an intention) and doing something intentionally are not the same phenomenon and that the former is more concerned with the coordination of an agent's plans. We agree, and in this paper we concentrate primarily on future-directed intentions. Hereafter, the term "intention" will be used in that sense only.

Intention has often been analyzed differently from other mental states such as belief and knowledge. First, whereas the content of beliefs and knowledge is usually considered to be in the form of propositions, the content of an intention is typically regarded as an action. For example, Castañeda (1975) treats the content of an intention as a "practition," akin (in computer science terms) to an action description. It is claimed that by doing so, and by strictly separating the logic of propositions from the logic of practitions, one avoids undesirable properties in the logic of intention, such as the fact that if one intends to do an action a, one must also intend to do a or b. However, it has also been argued that needed connections between propositions and practitions may not be derivable (Bratman 1983).

Searle (1983) claims that the content of an intention is a causally self-referential representation of its conditions of satisfaction (and see also Harman 1986). That is, for an agent to intend to go to the store, the conditions of satisfaction would be that the intention should cause the agent to go to the store. Our analysis is incomplete in that it does not deal with this causal self-reference. Nevertheless, the present analysis will char-

acterize many important properties of intention discussed in the philosophical literature.

A second difference among kinds of propositional attitudes is that some, such as belief, can be analyzed in isolation—one axiomatizes the properties of belief apart from those of other attitudes. However, intention is intimately connected with other attitudes, especially belief, as well as with time and action. Thus, any formal analysis of intention must explicate these relationships. In the next section we explore what it is that theories of intention should handle.

### 1.3 *Desiderata for a Theory of Intention*

Bratman (1987) argues that rational behavior cannot just be analyzed in terms of beliefs and desires (as many philosophers have held). A third mental state, intention, which is related in many interesting ways to beliefs and desires but is not reducible to them, is necessary. There are two justifications for this claim. First, noting that agents are resource-bounded, Bratman suggests that no agent can continually weigh his[4] competing desires, and concomitant beliefs, in deciding what to do next. At some point the agent must just *settle on* one state of affairs for which to aim. Deciding what to do establishes a limited form of *commitment*. We will explore the consequences of such commitments.

A second reason is the need to coordinate one's future actions. Once a future act is settled on—that is, intended—one typically decides on other future actions to take with that action as given. This ability to plan to do some act $A$ in the future, and to base decisions on what to do subsequent to $A$, requires that a rational agent *not* simultaneously believe he will *not* do $A$. If he did, the rational agent would not be able to plan past $A$ since he believes it will not be done. Without some notion of commitment, deciding what else to do would be a hopeless task.

Bratman argues that unlike mere desires, intentions play the following three functional roles:

1. *Intentions normally pose problems for the agent; the agent needs to determine a way to achieve them.* For example, if an agent intends to fly to New York on a certain date and takes on actions to enable himself to do so, then the intention affected the agent in the right way.
2. *Intentions provide a "screen of admissibility" for adopting other intentions.* Whereas desires can be inconsistent, agents do not normally adopt intentions that they believe conflict with their present- and future-directed intentions. For example, if an agent intends to hardboil an egg and knows he has only one egg (and cannot get any more in time), he should not simultaneously intend to make an omelette.
3. *Agents "track" the success of their attempts to achieve their intentions.* Not only do agents care whether their attempts succeed, but they are disposed to replan to achieve the intended effects if earlier attempts fail.



In addition to the above functional roles, it has been argued that intending should satisfy the following properties. If an agent intends to achieve p, then:

4. *The agent believes* p *is possible.*
5. *The agent does not believe he will not bring about* p.[5]
6. *Under certain conditions, the agent believes he will bring about* p.
7. *Agents need not intend all the expected side effects of their intentions*[6].

For example, imagine a situation not too long ago in which an agent has a toothache. Although dreading the process, the agent decides that he needs desperately to get his tooth filled. Being uninformed about anaesthetics, the agent believes that the process of having his tooth filled will necessarily cause him much pain. Although the agent intends to ask the dentist to fill his tooth, and, believing what he does, he is willing to put up with pain, the agent could surely deny that he thereby intends to be in pain.

Bratman argues that what one intends is, loosely speaking, a subset of what one chooses. Consider an agent as choosing one desire to pursue from among his competing desires and, in so doing, choosing to achieve some state of affairs. If the agent believes his action(s) will have certain effects, the agent has chosen those effects as well. That is, one chooses a "scenario" or a possible world. However, one does not intend everything in that scenario; for example, one need not intend harmful expected side effects of one's actions (though if one knowingly brings them about as a consequence of one's intended action, they have been brought about *intentionally*). Bratman argues that side effects do not play the same roles in the agent's planning as true intentions do. In particular, they are not goals whose achievement the agent will track; if the agent does not achieve them, he will not go back and try again.

We will develop a theory in which expected side effects are *chosen* but not intended.

### 1.4 *Intentions as a Composite Concept*

Intention will be modeled as a composite concept specifying what the agent has chosen and how the agent is committed to that choice. First, consider the desire that the agent has chosen to pursue as put into a new category. Call this chosen desire, loosely, a goal.[7] By construction, chosen desires are consistent. We will give them a possible-worlds semantics, and

hence the agent will have chosen a set of worlds in which the goal/desire holds.

Next, consider an agent to have a *persistent goal* if he has a goal (that is, a chosen set of possible worlds) that will be kept at least as long as certain conditions hold. For example, for a fanatic these conditions might be that his goal has not been achieved but is still achievable. If either of those circumstances fails, even the fanatical agent must drop his commitment to achieving the goal. Persistence involves an agent's *internal* commitment to a course of events over time.[8] Although a persistent goal is a composite concept, it models a distinctive state of mind in which agents have both chosen and committed to a state of affairs.

We will model intention as a kind of persistent goal. This concept, and especially its variations allowing for subgoals, interpersonal subgoals, and commitments relative to certain other conditions, is interesting for its ability to model much of Bratman's analysis. For example, the analysis shows that agents need not intend the expected side effects of their intentions because agents need not be committed to the expected consequences of those intentions. To preview the analysis, persistence need not hold for expected side effects because the agent's *beliefs* about the linkage of the act and those effects could change.

Strictly speaking, the formalism predicts that agents only intend the logical equivalences of their intentions, and in some cases intend their logical consequences. Thus, even using a possible-worlds approach, one can get a fine-grained modal operator that satisfies many desirable properties of a model of intention.

## 2 *Methodology*

### *2.1 Strategy: A Tiered Formalism*

The formalism will be developed in two layers: atomic and molecular. The foundational atomic layer provides the primitives for the rational action. At this level can be found the analysis of beliefs, goals, and actions. Most of the work here is to sort out the relationships among the basic modal operators. Although the primitives chosen are motivated by the phenomena to be explained, few commitments are made at this level to details of theories of rational action. In fact, many theories could be developed with the same set of primitive concepts. Thus, at the foundational level we provide a framework in which to express such theories.

The second layer provides new concepts defined out of the primitives. Upon these concepts, we develop a partial theory of rational action. Defined concepts provide economy of expression and may themselves be of theoretical significance because the theorist has chosen to form some definitions and not others. The use of defined concepts elucidates the origin of their important properties. For example, in modeling intention with persistent goals, one can see how various properties depend on particular primitive concepts.



Finally, although we do not do so in this paper (but see chapter 12), one can erect theories of rational interaction and communication on this foundation. By doing so, properties of communicative acts can be derived from the embedding logic of rational interaction, whose properties are themselves grounded in rational action.

### *2.2 Successive Approximations*

The approach to be followed in this paper is to approximate the needed concepts with sufficient precision to enable us to explore their interactions. We do not take as our goal the development of an exceptionless theory but rather will be content to give plausible accounts that cover the important and frequent cases. Marginal cases (and arguments based on them) will be ignored when developing the first version of the theory.

### *2.3 Idealizations*

The research presented here is founded on various idealizations of rational behavior. Just as initial progress in the study of mechanics was made by assuming frictionless planes, so too can progress be made in the study of rational action with the right idealizations. Such assumptions should approximate reality—for example, beliefs can be wrong and revised, goals not achieved and dropped—but not so closely as to overwhelm. Ultimately, choosing the right initial idealizations is a matter of research strategy and taste.

A key idealization we make is that no agent will attempt to achieve something forever—everyone has limited persistence. Similarly, agents will be assumed not to procrastinate forever. Although agents may adopt commitments that can only be given up when certain conditions ($C$) hold, the assumption of limited persistence requires that the agent eventually drop each commitment. Hence, it can be concluded that eventually conditions $C$ hold. Only because of this assumption are we able to draw conclusions from an agent's adopting a persistent goal. Our strategy will be first to explore the consequences of fanatical persistence—commitment to a goal until it is believed to be achieved or unachievable. Then, we will weaken the persistence conditions to something more reasonable.

### *2.4 Map of the Paper*

In the next sections of the paper we develop elements of a formal theory of rational action, leading up to a discussion of persistent goals and the consequences that can be drawn from them with the assumption of limited

persistence. Then, we demonstrate the extent to which the analysis satisfies the above-mentioned desiderata for intention and show how the analysis of intention solves various classical problems. Finally, we extend the underlying concept of a persistent goal to a more general one and briefly illustrate the utility of that more general concept for rational interaction and communication. In particular, we show how agents can have interlocking commitments.

## 3 *Elements of a Formal Theory of Rational Action*

The basis of our approach is a theory of rational action. The theory is expressed in a logic whose model theory is based on a possible-worlds semantics. We propose a logic with four primary modal operators: BELief, GOAL, HAPPENS (what event happens next), and DONE (which event has just occurred). With these operators, we will characterize what agents need to know to perform actions that are intended to achieve their goals. The world will be modeled as a linear sequence of events (similar to linear time temporal models; see Lamport 1980, Lansky 1985).[9] By adding GOAL, we can model an agent's intentions.

As a general strategy, the formalism will be too strong. First, we have the usual consequential closure problems that plague possible world models for belief. These, however, will be accepted for the time being, and we welcome attempts to develop finer-grained semantics (for instance, Barwise and Perry 1983; Fagin and Halpern 1985). Second, the formalism will describe agents as satisfying certain properties that might generally be true, but for which there might be exceptions. Perhaps a process of nonmonotonic reasoning could smooth over the exceptions, but we will not attempt to specify such reasoning here (but see Perrault, this volume). Instead, we assemble a set of basic principles and examine their consequences for rational interaction. Finally, the formalism should be regarded as a description or specification *of* an agent, rather than one that any agent could or should use.

Most of the advantage of the formalism stems from the assumption that agents have a limited tolerance for frustration; they will not work forever to achieve their goals. Yet, because agents are (often) persistent in achieving their goals, they will work to achieve them. Hence, although all goals will be dropped, they will not be droped too soon.

### 3.1 *Syntax*

The language we will use has the usual connectives of a first-order language with equality, as well as operators for the propositional attitudes and for talking about sequences of events:



- (BEL x p) and (GOAL x p), which say that x has p as a belief or goal, respectively;
- (AGT x e), which that x is the only agent of the sequence of events e;
- $(e_1 \leq e_2)$, which says that $e_1$ is an initial subsequence of $e_2$;
- time propositions, which say that it is a certain time, as described below; and
- (HAPPENS a) and (DONE a), which say that a sequence of events describable by an *action expression* a will happen next or just happened, respectively.

An action expression here is built from variables ranging over sequences of events using the constructs of dynamic logic (see, for example, Harel 1979; Moore 1980; Pratt 1978):

- a;b is action composition;
- a|b is nondeterministic choice;
- p? is a test action;
- a* is repetition.

The usual programming constructs like IF/THEN actions and WHILE loops will be formed from these (see section 3.2.1). We will use e as a variable ranging over sequences of events, and a and b for action expressions.

Note that there are no syntactic constructs for (what we are calling) persistent goals or intentions. These will be defined out of the above formulas. For simplicity, we adopt a logic with no singular terms, using instead predicates and existential quantifiers. However, for readability, we will sometimes use constants. The interested reader can expand these into the full predicative form if desired.

### 3.2 *Semantics*

We will adapt the usual possible-worlds model for belief to goals and events. Informally, a possible world is a string of events that temporally extends infinitely in the past and future and characterizes a possible way the world could have been and could be. Because things will naturally change over a course of events, the truth of a proposition in our language depends not only on world in question but also on an index into that course of events (roughly, a time point).

For the sake of simplicity, these indices are modeled as integers, and possible worlds are modeled by elements of a set, $T$, of functions from the integers into a set, $E$, of primitive event types. If $\sigma \in T$, then $\sigma(n)$ is understood as the (unique) event that is happening at point $n$. We also assume that each event type has a single agent (taken from a set, $P$, of people) as given by a function $Agt$. For this paper, the only events that will be

considered are those performed by an agent. Although there are no simultaneous primitive events in this model, an agent is not guaranteed to execute a sequence of events without events performed by other agents intervening.

To handle the part of the language that derives from standard first-order logic, we need a domain of quantification $D$ that includes all people and finite sequences of events, and a relation $I$, which at every world and index point assigns to each $k$-place predicate symbol a $k$-ary relation over $D$. The truth of atomic sentences, conjunctions, quantifications, and so forth, is determined in the usual way. A formula (AGT x e) is true when the person designated by x is the agent of every event in the sequence designated by e. A formula $e_1 \leq e_2$ is true when the event sequence designated by $e_1$ is an initial subsequence of that designated by $e_2$.

Time propositions are currently just numerals that can be used to name index points on a course of events. However, for ease of exposition, we will write them as if they were time-date expressions such as 2:30PM/3/6/85. These will be true or false in a world at a given index iff the index is the same as that denoted by the time proposition (that is, numeral). Depending on the problem at hand, we may use timeless propositions, such as (At Robot NY). Other problems are more accurately modeled by conjoining a time proposition, such as (At Robot NY) $\wedge$ 2:30PM/3/6/85. Thus, if the above conjunction were a goal, both conjuncts would have to be true simultaneously.

Turning now to the attitudes, we imagine that at any given point agents have a collection of possibly conflicting or even contradictory opinions and desires about the world. We assume, however, that when it comes time to act, each agent chooses (if only momentarily) a consistent set of each to work with.[10] We call these the beliefs and goals, respectively, of the agent. As the world evolves—that is, agents perform actions—their beliefs and goals can and will normally change.

Although each world has a fixed, predetermined future, agents usually do not know which world they are in. Instead, some of these worlds are compatible with the agent's beliefs and goals. This is specified by means of two accessibility relations $B$ and $G$. For a given agent $x$, $B(\varphi, x, n, \sigma^*)$ holds if the world $\sigma^*$ could be the actual one as far as $x$ is concerned at point $n$ (that is, $x$ has no beliefs at that point that could rule it out). Similarly, $G(\sigma, x, n, \sigma^*)$ holds if the world $\sigma^*$ is satisfactory as far as $x$ is concerned at point $n$ (that is, $x$ has no goals at that point that could rule it out). Stated differently, the $B$ relation holds if $\sigma^*$ is compatible with what $x$ believes in world $\varphi$ at point $n$ (and similarly for $G$ and goals). Thus, a formula (BEL x p) is true when p is true at every $B$-related world, and (GOAL x p) is true when p is true at every $G$-related world.

Turning now to actions, we must first be clear about what it means for an action to occur between two index points: e occurs through an interval



if it designates the event sequence in that interval; a;b occurs if a occurs from the start up to some intermediate point and then b occurs from that point to the final one; a|b occurs if either a or b does; the test action, p?, occurs between two index points if they are the same point (that is, no events are involved) and if at that point p is true; finally, a* occurs through an interval if the interval can be broken into some number of subintervals and a occurs through each of them (in other words, a* means the repetition of a some number of times).

Given this, a formula (HAPPENS a) is true at some index point on a world when there is a subsequent (future) index point on that world such that a describes the sequence of events between the two index points. Similarly, (DONE a) is true at some index point when there is a previous (past) index point such that a describes the sequence of events between the two index points. So the former says that something describable by a will happen next, whereas the latter says that it just happened.

This completes our informal description of truth for our language. A semantic structure $M$ has as components $D$, $P$, $E$, $Agt$, $T$, $B$, $G$, $I$, as described above. However, we will impose various additional constraints and assumptions on these components as we progress. A formula p is *satisfiable* if there is at least one structure $M$ (satisfying these constraints and assumptions), world $\sigma \in T$, and integer index $n$, such that p is true at $\sigma$ and $n$ for this $M$. A well-formed formula p is *valid*, written $\models$ p, if its negation is not satisfiable.

*3.2.1 Abbreviations* It will be convenient to adopt the following abbreviations:

*Empty sequence*: nil $\stackrel{def}{=}$ (∀x (x=x))?. a=NIL $\stackrel{def}{=}$ ∀b (a≤b).
As a test action, NIL always succeeds; as an event sequence, it is a subsequence of every other one.

*Conditional action*: [IF p THEN a ELSE b] $\stackrel{def}{=}$ p?;a | ~p?;b.
That is, as in dynamic logic, an if-then-else action is a disjunctive action of doing action a at a time at which p is true or doing action b at a time at which p is false. Note that the semantics of a conditional action does not require that the condition be believed by someone to be true. However, when agents execute conditionals with disjoint branches, they will have to believe the condition is true (or believe it is false).

*While-loops*: [WHILE p DO a] $\stackrel{def}{=}$ (p?;a)*; ~p?
While-loops are a sequence of doing action a zero or more times, prior to each of which p is true. After the iterated action stops, p is false.

*Eventually*: $\Diamond p \stackrel{def}{=} \exists x$ (HAPPENS x;p?).
In other words, $\Diamond p$ is true (in a given possible world) if there is something that happens after which p holds, that is, if p is true at some point in the future.
*Always*: $\Box p \stackrel{def}{=} \sim \Diamond \sim p$.
$\Box p$ means that p is true throughout the course of events.
A useful application of $\Box$ is $\Box(p \supset q)$, in which no matter what happens, p still implies q. We can now distinguish between $p \supset q$'s being logically valid, its being true in all courses of events, and its merely being true after some event happens.

3.2.2 *Constraints on the Model* We impose the following constraints on the model:

*Consistency*: $B$ is Euclidean, transitive, and serial; $G$ is serial. $B$'s being Euclidean essentially means that the worlds the agent thinks are possible (given what is believed) form an equivalence relation but do not necessarily include the real world (Halpern and Moses 1985). Seriality implies that beliefs and goals are (separately) consistent. This is enforced by there always being a world that is either $B$- or $G$-related to a given world.
*Realism*: $\forall \sigma, \sigma^*$, if $\langle \sigma, n \rangle G[p]\sigma^*$, then $\langle \sigma, n \rangle B[p]\sigma^*$. In other words, $G \subseteq B$. That is, the worlds that are consistent with what the agent has chosen are not ruled out by his beliefs. Without this constraint, the agent could choose worlds involving (for example) future events that he believes will never happen. We believe this condition to be so strong, and its model-theoretical statement so simple, that it deserves to be imposed as a constraint. It ensures that an agent does not want the opposite of what he believes to be unchangeable. For example, assume an agent knows that he will die in two months (and he does not believe in life after death). One would not expect that agent, if still rational, to buy a plane ticket to Miami in order to play golf three months hence. Simply, an agent cannot choose such worlds since they are not compatible with what he believes.

3.3 *Properties of the Model*
We begin by exploring the temporal and action-related aspects of the model, describing properties of our modal operators HAPPENS, DONE, and $\Diamond$. Next, we discuss belief and relate it to the temporal modalities. Then, we explore the relationships among all these and GOAL. Finally, we characterize an agent's persistence in achieving a goal.

Valid properties of the model are termed "Propositions." Properties that constitute our theory of the interrelationships among agent's beliefs, goals, and actions will be stated as "Assumptions." These are essentially non-logical axioms that constrain the models that we consider.[11] The term "Theorem" is reserved for major results.



3.3.1 *Events and Action Expressions* The framework proposed here separates primitive events from action expressions. Examples of primitive events might include moving an arm, grasping, exerting force, and uttering a word or sentence. Action expressions denote sequences of primitive events that satisfy certain properties. For example, a movement of a finger may result in a circuit being closed, which may result in a light coming on. We will say that one primitive event happened, but one that can be characterized by various complex action expressions. This distinction between primitive events and complex action descriptions must be kept in mind when characterizing real-world phenomena or natural-language expressions.

For example, to say that an action a occurs, we use (HAPPENS a). To characterize world states that are brought about, we use (HAPPENS $\sim$p?;a;p?), saying that event a brings about p. To be a bit more concrete, one would not typically have a primitive event type for closing a circuit. So, to say that John closed the circuit, one would say that John did something (perhaps a sequence of primitive events) causing the circuit to be closed—$\exists$e (DONE $\sim$(Closed c)?;e;(Closed c)?.

Another way to characterize actions and events is to have predicates be true of them. For example, one could have (Walk e) to mean that a given event (type) is a walking event. This way of describing events has the advantage of allowing complex properties (such as running a race) to hold for an undetermined (and unnamed) sequence of events. However, because the predications are made about the events, not the attendant circumstances, this method does not allow us to describe events performed only in certain circumstances. We will need to use both methods for describing actions.

3.3.2 *Properties of Acts/Events under* HAPPENS We adopt the usual axioms characterizing how complex action expressions behave under HAPPENS, as treated in a dynamic logic (for example, Harel 1979; Moore 1980; Pratt 1978), including the following:

*Proposition 1* Properties of complex acts

$\models$ (HAPPENS a;b) $\equiv$ (HAPPENS a;(HAPPENS b)?).

$\models$ (HAPPENS a|b) $\equiv$ (HAPPENS a) V (HAPPENS b).

$\models$ (HAPPENS p?;q?) $\equiv$ p $\wedge$ q.

$\models$ (HAPPENS a*;b) $\equiv$ (HAPPENS b | a;a*;b).

That is, action a;b happens next iff a happens next, producing a world state in which b then happens next. The "nondeterministic choice" action a|b (read "|" as "or") happens next iff a happens next or b does. The text action action p? happens next iff p is currently true. Finally, the iterative action a*;b happens next iff b happens or one step of the iteration has been taken, followed by a*;b again.

Among many additional properties, note that after doing action a, a would have just been done:

*Proposition 2*
$\models$ (HAPPENS a) $\equiv$ (HAPPENS a;(DONE a)?).

Also, if a has just been done, then just prior to its occurence, it was going to happen next:

*Proposition 3*
$\models$ (DONE a) $\equiv$ (DONE (HAPPENS a)?;a).

Although this may seem to say that the unfolding of the world is determined only by what has just happened, and is not random, this determinacy is entirely moot for our purposes. Agents need never know what possible world they are in and hence what will happen next. More serious would be a claim that agents have no "free will"—what happens next is determined without regard to their intentions. However, as we will see, this is not a property of agents; their intentions constrain their future actions. Next, observe that a test action is done whenever the condition holds:

*Proposition 4*
$\models$ p $\equiv$ (DONE p?).

That is, the test action filters out courses of events in which the proposition tested is false. The truth of Proposition 4 follows immediately from the definition of "?".

For convenience, let us define versions of DONE and HAPPENS that specify the agent of the act:

*Definition 1*
(DONE x a) $\stackrel{def}{=}$ (DONE a) $\wedge$ (AGT x a).

*Definition 2*
(HAPPENS x a) $\stackrel{def}{=}$ (HAPPENS a) $\wedge$ (AGT x a).

Finally, one distinction is worth pointing out. When action variables are bound by quantifiers, they range over sequences of events (more precisely, event types). When they are left free in a formula, they are intended as schematic and can be instantiated with complex action expressions.



*3.3.3 Temporal Modalities:* DONE, $\Diamond$, and $\Box$ Temporal concepts are introduced with DONE (for past happenings) and $\Diamond$ (read "eventually"). To say that p was true at some point in the past, we use $\exists$e (DONE p?;e). $\Diamond$ is to be regarded in the "linear time" sense and is defined above. Essentially, $\Diamond$p is true iff somewhere in the future p becomes true. $\Diamond$p and $\Diamond \sim$p are jointly satisfiable. Since $\Diamond$p starts "now," the following property is also true:

*Proposition 5*
$\models$ p $\supset \Diamond$p.

The following are trivial consequences:

*Proposition 6*
$\models \Diamond(p \vee q) \wedge \Box \sim q \supset \Diamond p$.

*Proposition 7*
$\models \Box(p \supset q) \wedge \Diamond p \supset \Diamond q$.

To talk about propositions that are not true now but will become true, we define:

*Definition 3*
(LATER p) $\stackrel{def}{=} \sim p \wedge \Diamond p$.

A property of this definition that follows from the equivalence of $\Diamond$p and $\Diamond\Diamond$p is:

*Proposition 8*
$\models \sim$(LATER $\Diamond$p).

*3.3.4 Constraining Courses of Events* We will have occasion to state constraints on courses of events. To do so, we define the following:

*Definition 4*
(BEFORE p q) $\stackrel{def}{=} \forall c$ (HAPPENS c;q?) $\supset \exists a$ (a $\leq$ c) $\wedge$ (HAPPENS a;p?).

This definition states that p comes before q (starting at index *n* in the course of events) if, whenever q is true in a course of events, p has been true (after the index *n*). Obviously,

*Proposition 9*
$\models \Diamond q \wedge$ (BEFORE p q) $\supset \Diamond p$.

That is, if q is eventually true, and q's being true requires that p has been true, then eventually p holds. Furthermore, we have:

*Proposition 10*
$\models \sim p \supset$ (BEFORE ($\exists$e (DONE $\sim$p?;e;p?)) p).

This basically says that worlds are consistent—no proposition changes truth-value without some event happening. In particular, there is no notion in this model for the simple passage of time (without any intervening events) affecting anyone's beliefs or goals. One would like to adopt the view that some event must *cause* that change, but as yet there is no primitive relation of causality.

### 3.4 *The Attitudes*

BEL and GOAL characterize what is *implicit* in an agent's beliefs and goals (chosen desires), rather than what an agent actively or explicitly believes, or has as a goal.[12] That is, these operators characterize what *the world would be like* if the agent's beliefs and goals were true. Importantly, we do not include an operator for wanting, since desires need not consistent. Although desires certainly play an important role in determining goals and intentions, we assume that once an agent has sorted out his possibly inconsistent desires in deciding what he wishes to achieve, the worlds he will be striving for are consistent.

*3.4.1 Belief* For simplicity, we assume the usual Hintikka-style axiom schemata for BEL (Halpern and Moses 1985) (corresponding to a "Weak S5" modal logic):

*Proposition 11* Belief axioms

a. $\models \forall x\,(\mathrm{BEL}\ x\ p) \wedge (\mathrm{BEL}\ x\ (p \supset q)) \supset (\mathrm{BEL}\ x\ q)$

b. $\models \forall x\,(\mathrm{BEL}\ x\ p) \supset (\mathrm{BEL}\ x\ (\mathrm{BEL}\ x\ p))$

c. $\models \forall x \sim(\mathrm{BEL}\ x\ p) \supset (\mathrm{BEL}\ x \sim(\mathrm{BEL}\ x\ p))$

d. $\models \forall x\,(\mathrm{BEL}\ x\ p) \supset \sim(\mathrm{BEL}\ x \sim p)$.

And we have the usual "necessitation" rule:

*Proposition 12*
*If* $\models p$ *then* $\models (\mathrm{BEL}\ x\ p)$.

If p is a theorem (in other words, is valid), then it follows from the agent's beliefs at all times. For example, all tautologies follow from the agent's beliefs. Clearly, we also have:

*Proposition 13*
*If* $\models p$ *then* $\models (\mathrm{BEL}\ x\ \Box p)$.

That is, theorems are believed to be always true. Also, we introduce KNOW by definition:

*Definition 5*
$(\mathrm{KNOW}\ x\ p) \stackrel{def}{=} p \wedge (\mathrm{BEL}\ x\ p)$.



Of course, this characterization of knowledge has many known difficulties, but it will suffice for present purposes. Next, we will say an agent is COMPETENT with respect to p if he is correct whenever he thinks p is true:

*Definition 6*
$(\mathrm{COMPETENT}\ x\ p) \stackrel{def}{=} (\mathrm{BEL}\ x\ p) \supset (\mathrm{KNOW}\ x\ p)$.

Agents competent with respect to some proposition p adopt only beliefs about that proposition for which they have good evidence. For the purposes of this paper, we assume that agents are competent with respect to the primitive actions they have done:

*Assumption 1*
$\models \forall x,e\,(\mathrm{AGT}\ x\ e) \supset [(\mathrm{DONE}\ e) \equiv (\mathrm{BEL}\ x\ (\mathrm{DONE}\ e))]$.

Note that this assumption does *not* hold when e is replaced by an arbitrary action expression, even if x is the agent. For example, if the agent does not know the truth-value of p after just doing a, the agent may have done the action a;p? without realizing it was done. But the assumption rules out unknowing execution of *primitive actions by an agent.*

We also assume that an agent cannot believe he is the agent of what is to happen next without his knowing what the next primitive event will be. That is, if an agent thinks he is about to do *something*, then there must be some initial sequence that he believes he is going to do next:

*Assumption 2*
$\models (\mathrm{BEL}\ x\ \exists e \neq \mathrm{NIL}\ (\mathrm{HAPPENS}\ x\ e)) \supset \exists e \neq \mathrm{NIL}\ (\mathrm{BEL}\ x\ (\mathrm{HAPPENS}\ x\ e))$.

The antecedent would be true if the agent believes he is about to do a complex action (for instance, one containing a disjunction, or an iteration until a condition is satisfied). So, there may be uncertainty in his mind about what he is about to do. But for anything to happen at all, we assume that there must be some nonempty initial sequence that he has settled on and thinks he is going to do next.

### 3.5 *Goals*

At a given point in a course of events, agents choose worlds they would like (most) to be in—ones in which their *goals* are true. (GOAL x p) is meant to be read as p is true in all worlds, accessible from the current world, that are compatible with the agent's goals. Roughly, p follows from the agent's goals. Since agents choose entire worlds, they choose the (logically and physically) necessary consequences of their goals. At first glance, this appears troublesome if we interpret the facts that are true in all worlds compatible with an agent's goals as intended. However, intention will involve a form of commitment that will rule out such consequences as being intended, although they are chosen.

GOAL has the following properties:

*Proposition 14* Consistency
$\models \forall x\,(\text{GOAL } x\ p) \supset \sim(\text{GOAL } x \sim p)$.

What is implicit in someone's goals is closed under consequence:

*Proposition 15*
$\models (\text{GOAL } x\ p) \wedge (\text{GOAL } x\ p \supset q) \supset (\text{GOAL } x\ q)$.

Again, we have a necessitation property:

*Proposition 16*
*If* $\models p$ *then* $\models (\text{GOAL } x\ p)$.

That is, if p is a theorem, it is true in all chosen worlds. However, agents can distinguish such "trivial" goals from others, as explained below.

*3.5.1 Achievement Goals* Agents can distinguish between achievement goals and maintenance goals. Achievement goals are those the agent believes to be currently false; maintenance goals are those the agent already believes to be true. We will not be concerned in this paper with maintenance goals. However, to characterize achievement goals, we use:

*Definition 7*
$(\text{A-GOAL } x\ p) \stackrel{def}{=} (\text{GOAL } x\ (\text{LATER } p)) \wedge (\text{BEL } x \sim p)$.

That is, x believes (and therefore accepts) that p is currently false, but in his chosen worlds p is eventually true. In other words, this is the more standard notion of goals, where what is desired for the future is something that is believed to be currently false.

*3.5.2 No Persistence/Deferral Forever* Agents are limited in both their persistence and their procrastination. They cannot try forever to achieve their goals; eventually they give up. On the other hand, agents do not forever defer working on their goals. The assumption below captures both of these desiderata:

*Assumption 3*
$\models \Diamond \sim(\text{GOAL } x\ (\text{LATER } p))$.

Thus, agents eventually drop all achievement goals. Because one cannot conclude that agents always act on their goals, one needs to guard against infinite procrastination. However, one could have an agent who forever fails to achieve his goals but believes success is still achievable. The limiting case here is an agent who executes an infinite loop. Another case is that of a compulsive gambler who continually thinks success is just around the corner. Our assumption rules out these pathological cases from consideration but still allows agents to try hard. Finally, since no one ever said the world is fair (in the computer science sense), an agent who is ready to act in what he believes to be the correct circumstance may never get a chance to execute his action because the world keeps changing. We only require that if faced with such monumental unfairness, the agent reach the conclusion that the act is impossible.



One might object that there are still achievement goals that agents could keep forever. For example, one might argue that the goal expressed by "I always want more money than I have" is kept forever (or at least as long as the agent is alive);[13] but consider a plausible logical representation of that sentence in our formal language:

$$\Box[\text{GOAL I } \exists x,y\,(\text{HAVE I } x) \wedge y > x \wedge (\text{LATER } (\text{HAVE I } y))].$$

This sentence may be true, but it does not express an achievement goal since at some points the existential part may be believed to be true (and the goal is merely to maintain that truth). To express the achievement aspect, it is necessary to quantify into the GOAL clause as in

$$\Box[\forall x\,(\text{KNOW I } (\text{HAVE I } x)) \supset (\text{A-GOAL I } \exists y\,(y > x \wedge (\text{HAVE I } y)))].$$

But here, there is no single sentence that the agent always has as a goal; the goal changes because of the quantified variables. Hence, one cannot argue he keeps anything as an achievement goal forever. Instead, the agent forever gets new achievement goals.

Important consequences will follow from Assumption 3 when combined with an agent's commitments. First, we need to examine what, in general, are the consequences of having goals.

*3.5.3 Goals and Their Consequences* Unlike BEL, GOAL needs to be characterized in terms of all the other modalities. In particular, we need to specify how goals interact with an agent's beliefs about the future.

The semantics of GOAL specifies that worlds compatible with an agent's goals must be included in those compatible with his beliefs. This is reflected in the following property:

*Proposition 17*
$\models (\text{BEL } x\ p) \supset (\text{GOAL } x\ p)$.

From the semantics of BEL and GOAL, one sees that p will be evaluated at the same point in the *B*- and *G*-accessible worlds. So, if an agent believes p is true now, he cannot now want it to be currently false; agents do not choose what they cannot change. Conversely, if p is now true in all the agent's chosen worlds, then the agent does not believe it is currently false. For example, if an agent believes he has not just done event e, then he

cannot have (DONE x e) as a goal. Of course, he *can* have (LATER (DONE x e)) as a goal.

This relationship between BEL and GOAL makes more sense when one considers the future. Let p be of the form $\Diamond$q. From Proposition 17, we derive that if the agent wants q to be true sometime in the future, he does not believe it will be forever false. Conversely, let p be a proposition of the form $\Box$q. So, if an agent believes q is forever true (an example would be a tautology), Proposition 17 says that any worlds that the agent chooses must have q's being true as well.

Notice that although an agent may have to put up with what he believes is inevitable, he may do so reluctantly, knowing that if he should change his mind about the inevitability of that state of affairs, his choices would change. For example, the following is satisfiable:

$$(\text{BEL x } \Diamond p \wedge \Box[\sim(\text{BEL x } \Diamond p) \supset (\text{GOAL x } \Box \sim p)]),$$

That is, the agent can believe p is inevitable (and hence in all the agent's chosen worlds p will eventually be true) but at the same time believe that if he ever stops believing it is inevitable, he will choose worlds in which it is never true.

Notice also that, as a corollary of Proposition 17, an agent's beliefs and goals "line up" with respect to his own primitive actions that happen next:

*Proposition 18*
$\models \forall x,e\ (\text{BEL x } (\text{HAPPENS x e})) \supset (\text{GOAL x } (\text{HAPPENS x e}))$.

That is, if an agent believes he is about to do something next, then its happening next is true in all his chosen worlds. Of course, "successful" agents are ones who choose what they are going to do before believing they are going to do it; they come to believe they are going to do something because they have made certain choices. We discuss this further in our treatment of intention.

Next, as another simple subcase, consider the *consequences* of facts the agent believes hold in all of that agent's chosen worlds:

*Proposition 19* Expected consequences
$\models (\text{GOAL x p}) \wedge (\text{BEL x } p \supset q) \supset (\text{GOAL x q})$.

By Proposition 17, if an agent believes $p \supset q$ is true, $p \supset q$ is true in all his chosen worlds. Hence, by Proposition 15, q follows from his goals as well.

At this point we are finished with the foundational level, having described agents' beliefs and goals, events, and time. In so doing, we have characterized agents as not striving for the unachievable and as eventually foregoing the contingent. What is missing is *commitment*, to ensure that none of these goals are given up too easily.



## 4 *Persistent Goals*

To capture *one* grade of commitment (fanatical) that an agent might have toward his goals, we define a persistent goal, P-GOAL, to be one that the agent will not give up until he thinks it has been satisfied, or until he thinks it will never be true. The latter case could arise easily if the proposition p is one that specifically mentions a time. Once the agent believes that time is past, he believes the proposition is impossible to achieve. Specifically, we have

*Definition 8*
$$(\text{P-GOAL x p}) \stackrel{def}{=} (\text{GOAL x } (\text{LATER p})) \wedge (\text{BEL x } \sim p) \wedge [\text{BEFORE } ((\text{BEL x p}) \vee (\text{BEL x } \Box \sim p)) \sim(\text{GOAL x } (\text{LATER p}))].$$

Notice the use of LATER, and hence $\Diamond$, above. Clearly, P-GOALs are achievement goals; the agent's goal is that p be true in the future, and he believes it is not currently true. As soon as the agent believes it will never be true, we know the agent must drop his goal (by Proposition 17), and hence his persistent goal. Moreover, as soon as an agent believes p is true, the belief conjunct of P-GOAL requires that he drop the persistent goal of achieving p. Thus, these conditions are necessary and sufficient for dropping a persistent goal. However, the BEFORE conjunct does *not* say that an agent *must* give up his *simple* goal when he thinks it is satisfied, since agents may have goals of maintenance. Thus, achieving one's persistent goals may convert them into maintenance goals.

### 4.1 *The Logic of* P-GOAL

The logic of P-GOAL is weaker than one might expect. Unlike GOAL, P-GOAL does not distribute over conjunction or disjunction, and it is closed only under logical equivalence. First, we examine conjunction and disjunction Then, we turn to implication.

*4.1.1 Conjunction, Disjunction, and Negation* P-GOAL behaves as follows under conjunction, disjunction, and negation:

*Proposition 20* The logic of P-GOAL

a. $\not\models (\text{P-GOAL x } p \wedge q) \supseteq (\text{P-GOAL x p}) \wedge (\text{P-GOAL x q})$.

b. $\not\models (\text{P-GOAL x } p \vee q) \supseteq (\text{P-GOAL x p}) \vee (\text{P-GOAL x q})$.

c. $\models (\text{P-GOAL x } \sim p) \supset \sim(\text{P-GOAL x p})$.

First, (P-GOAL x $p \wedge q$) does not imply (P-GOAL x p) $\wedge$ (P-GOAL x q) because, although the antecedent is true, the agent might believe q is already true and thus cannot have q as a P-GOAL.[14] Conversely, (P-GOAL

x p) ∧ (P-GOAL x q) does not imply (P-GOAL x p∧q), because (GOAL x (LATER p)) ∧ (GOAL x (LATER q)) does not imply (GOAL x (LATER p∧q)); p and q could be true at different times.

Similarly, (P-GOAL x p∨q) does not imply (P-GOAL x p) ∨ (P-GOAL x q) because (GOAL x (LATER p∨q)) does not imply (GOAL x (LATER p)) ∨ (GOAL x (LATER q)); p could come to hold in some possible worlds compatible with the agent's goals, and q in others. However, neither p nor q is forced to hold in all G-accessible worlds. Moreover, the implication does not hold in the other direction either, because of the belief conjunct of P-GOAL; although the agent may believe ∼p or he may believe ∼q, that does not guarantee he believes ∼(p∨q) (that is, ∼p∧q).

With respect to the last property, note that although it is impossible to be committed to achieving both p and ∼p (since one of them is not believed to be false), it is quite possible to be committed to achieving (p∧ ◇∼p). However, because of Proposition 8, which says that (LATER ◇p) is always false, (GOAL x (LATER ◇p)) is always false, and so (P-GOAL x ◇p) is always false.

*4.1.2 No Consequential Closure of* P-GOAL We demonstrate that P-GOAL is closed only under logical equivalence. Below are listed the possible relationships between a proposition p and a consequence q, which we term a *side effect*. Assume in all cases that (P-GOAL x p). Then, depending on the relationship of p to q, we have the cases shown in table 3.1. We will say a "case" fails, indicated by an "N" in the third column, if (P-GOAL x q) does not hold.

Case 1 fails for a number of reasons, most importantly because the agent's persistent goals depend on his beliefs, not on the facts. However, consider Case 2. Even though the agent may believe p⊃q holds, Case 2 fails because that implication cannot affect the agent's persistent goals, which refer to p's being true *later*. That is, the agent believes p is false and does not have the goal of its currently being true.

**Table 3.1**
P-GOAL and progressively stronger relationships between p and q.

| Case | Relationship of p to q | (P-GOAL x q)? |
|---|---|---|
| 1 | p ⊃ q | N |
| 2 | (BEL x (p ⊃ q)) | N |
| 3 | (BEL x □(p ⊃ q)) | N |
| 4 | □(BEL x □(p ⊃ q)) | Y/N |
| 5 | ⊨ p ⊃ q | Y/N |
| 6 | ⊨ p ≡ q | Y |



Consider Case 3, where the agent believes the implication *always* holds. Although Proposition 17 tells us that the agent has q as a goal, we show that the agent does not have q as a *persistent* goal. Recall that P-GOAL was defined so that the only reason an agent could give up a persistent goal was if it were believed to be satisfied or believed to be forever false. However, side effects are goals only because of a belief. If the belief changes, the agent need no longer choose worlds in which p⊃q holds and thus need no longer have q has a goal. However, the agent would have dropped the goal for reasons other than those stipulated by the definition of persistent goal and so does not have it as a persistent goal.

Now consider Case 4, in which the agent *always* believes the implication. Again, q need not be a persistent goal but for a different reason. Here, an agent could believe the side effect already held. Hence, by the second clause in the definition of P-GOAL, the agent would not have a persistent goal. This reason also blocks Case 5, closure under logical consequence. However, instances of Case 4 and Case 5 in which the agent does not believe the side effect already holds *would* require the agent to have the side effect as a persistent goal. Thus, we do not get closure in these cases, but because of what we believe to be the wrong reasons. A finer-grained semantic model than possible worlds might block closure in a more satisfying way by allowing agents to direct their goals toward situations that do not include side effects. Finally, in Case 6, where q is logically equivalent to p, the agent has q as a persistent goal. Having shown what cannot be deduced from P-GOAL, we now turn to its major consequences.

### *4.2 Persistent Goals Constrain Future Beliefs and Actions*

An important property of agents is that they eventually give up their achievement goals (Assumption 3). Hence, if an agent takes on a P-GOAL, he must give it up subject to the constraints imposed by P-GOAL:

*Proposition 21*
⊨ (P-GOAL x q) ⊃ ◇[(BEL x q) ∨ (BEL x □∼q)].

This proposition is a direct consequence of Assumption 3, the definition of P-GOAL, and Proposition 6. In other words, because agents eventually give up their achievement goals, and because the agent has adopted a persistent goal to bring about such a proposition q, eventually the agent must believe q or believe q will never come true. We now give a crucial theorem:

*Theorem 1* From persistence to eventualities
*If someone has a persistent goal of bringing about* p, p *is within his area of competence, and, before dropping his goal, the agent will not believe* p *will never occur, then eventually* p *becomes true:*

$$\models (\text{P-GOAL y p}) \wedge \Box(\text{COMPETENT y p}) \wedge \sim(\text{BEFORE (BEL y } \Box\sim p)\ \sim(\text{GOAL y (LATER p)})) \supset \Diamond p.$$

*Proof*
By Proposition 21, the agent eventually believes either that p is true or that p is unachievable. If he eventually thinks p is true, since he is always competent with respect to p, he is correct. The other alternative sanctioned by Proposition 21, that the agent believes p is unachievable, is ruled out by the assumption that (it so happens to be the case that) any belief of the agent that the goal is unachievable can come only *after* the agent drops his goal. Hence, by Proposition 6, the goal comes about. ■

If an agent who is not competent with respect to p adopts p as a persistent goal, we cannot conclude that eventually p will be true, since the agent could incorrectly come to believe that p holds. If the goal is not persistent, we also cannot conclude $\Diamond p$, since the agent could give up the goal without achieving it. If the goal actually is unachievable, but the agent does not know this and commits to achieving it, then we know that eventually, perhaps after trying hard to achieve it, the agent will come to believe it is forever false and give up.

*4.2.1 Acting on Persistent Goals* As mentioned earlier, one cannot conclude that, merely by committing to a chosen proposition (set of possible worlds), the agent will act; someone else could bring about the desired state of affairs. However, if the agent knows that he is the only one who could bring it about, then, under certain circumstances, we can conclude the agent will act. For example, propositions of the form (DONE x a) can only be brought about by the agent x. So, if an agent always believes the act a can be done (or at least believes it for as long as he keeps the persistent goal), the agent will act.

A simple instance of Proposition 21 is one where q is (HAPPENS x a). Such a goal is one where the next thing that happens is his doing action a. Eventually, either the agent believes the next action is his, or the agent eventually comes to believe he will never get the chance to perform it. We cannot guarantee that the agent will actually do the action next, for someone else could act before him. If the agent never believes his act will never be done, then by Proposition 21, the agent will eventually believe (HAPPENS x a). By Proposition 18, we know that (GOAL x (HAPPENS x a)). *If* the agent acts just when he believes the next act is his, we know that he did so believing it would happen next and having its happening next as his goal. One could say, loosely, that the agent acted "intentionally."



Bratman (1984) argues that one applies the term "intentionally" to foreseen consequences as well as to truly intended ones. That is, one intends a subset of what is done intentionally. Proposition 18 requires only that agents have expected effects as goals, not that they have them as persistent goals. Hence, the agent would in fact bring about intentionally all those foreseen consequences of his goal that actually obtain from his doing the act. However, he would not be committed to bringing about the side effects, and thus he did not intend to do so.

If agents adopt *time-limited* goals, such as (BEFORE (DONE x e) 2:30PM/6/24/86), one *cannot* conclude the agent definitely will act *in time*, even if he believes it is possible to act. Simply, the agent might wait too long. However, one *can* conclude (see below) that the agent will not adopt another persistent goal to do a non-NIL act he believes would make the persistent goal unachievable. Still, the agent could unknowingly (and hence, by Proposition 18, accidentally) make his persistent goal forever false. If one makes the further assumption that agents always know what they are going to do just before doing it, then one can conclude that agents will not in fact do anything to make their persistent goals unachievable.

All these conclusions are, we believe, reasonable. However, they do not indicate what the "normal" case is. Instead, we have characterized the possibilities, and we await a theory of default reasoning to further describe the situation.

One final complication worth noting is that even if we assume that agents are perfectly competent about beliefs and goals, it is unreasonable to assume that they are competent about their persistent goals. They may have incorrect beliefs about the BEFORE clause and misjudge the conditions under which they give up their achievement goals. A simple case is an agent who makes a promise (perhaps hastily), thinks he is committed, and then, finding out more about the situation, changes his mind and drops his goal without believing that it is satisfied or unachievable. Given that P-GOAL is based on whether an agent really is committed, the question remains as to the role of beliefs in one's commitments in a theory of this type.

## 5 *Intention as a Kind of Persistent Goal*

With our foundation laid, we are now in a position to define the concept of intention. There will be two defining forms for INTEND, depending on whether the argument is an action or a proposition.

### 5.1 $\text{INTEND}_1$

Typically, one intends to do actions. Accordingly, we define $\text{INTEND}_1$ to take an action expression as its argument.

*Definition 9*
$(\text{INTEND}_1\ x\ a) \stackrel{def}{=} (\text{P-GOAL}\ x\ [\text{DONE}\ x\ (\text{BEL}\ x\ (\text{HAPPENS}\ a))?;a])$,
*where* a *is any action expression.*

Let us examine what this says. First of all, (fanatically) intending to do an action a is a special kind of commitment (that is, persistent goal) to have done a. However, it is not a commitment just to doing a, for that would allow the agent to be committed to doing something accidentally or unknowingly. It seems reasonable to require that the agent be committed to believing he is about to do the intended action, and then doing it. Thus, intentions are future-directed, but here directed toward something happening *next*. This is as close as we can come to present-directed intention.

Second, it is a commitment to success—to having done the action. As a contrast, consider the following inadequate definition of $\text{INTEND}_1$:

$(\text{INTEND}_1\ x\ a) \stackrel{def?}{=} (\text{P-GOAL}\ x\ \exists e\ (\text{HAPPENS}\ x\ e;\ [\text{DONE}\ x\ a]?))$.

This would say that an intention is a commitment to being *on the verge of* doing some event e, after which x would have just done a.[15] Of course, being on the verge of doing something is not the same as doing it; any unforeseen obstacle could permanently derail the agent from ever performing the intended act. This would not be much of a commitment.

*5.1.1 Intending Actions* Let us apply $\text{INTEND}_1$ to each kind of action expression. First, consider intentions to "test" p. $(\text{INTEND}_1\ x\ p?)$ expands into

$(\text{P-GOAL}\ x\ (\text{DONE}\ x\ [\text{BEL}\ x\ (\text{HAPPENS}\ x\ p?)]?;p?))$.

By Proposition 1, this is equivalent to $(\text{P-GOAL}\ x\ (\text{DONE}\ x\ (\text{KNOW}\ x\ p)?))$, which reduces to $(\text{P-GOAL}\ x\ (\text{KNOW}\ x\ p))$. That is, the agent is committed to coming to know p (and he does not know it now). However, the agent is not committed to bringing about p himself.

Second, consider action expressions of the form e;p?. An example would be felling a tree:

$\exists e\ (\text{Chopping}\ e\ T) \wedge (\text{Tree}\ T) \wedge (\text{INTEND}_1\ x\ e;(\text{Down}\ T)?)$.

That is, there is a chopping event (type) e, such that the agent is committed to felling the tree by doing e, and he believes just prior to doing it that it that it will indeed fell the tree. Notice that e is quantified outside of the $\text{INTEND}_1$. This type of intention is appropriate when there is a fixed event (or event sequence) that an agent is willing to commit to. For example, with a small tree and a large axe, an agent may be very confident that the chopping event will do the trick.

However, not all trees are like this. Fortunately, chopping events can be repeated, although it need not be obvious how many times. Thus, certain intentions cannot be characterized in terms of a fixed sequence of events—an agent may never come to believe of any given event sequence that it will achieve the intention. In this case the intention might be expressed by

$\exists e\ (\text{Chopping}\ e\ T) \wedge (\text{Tree}\ T) \wedge$
$(\text{INTEND}_1\ x\ [\text{WHILE} \sim(\text{Down}\ T)\ \text{DO}\ e])$.

That is, the agent intends to do e repeatedly until the tree is down. It is important to notice that at no time does the agent need to know precisely which chopping event will finally knock down the tree. Instead, the agent is committed solely to executing the chopping event until the tree is down. To give up the commitment (that is, the persistent goal) constituting the intention, the agent must eventually come to believe he has done the iterative action believing it was about to happen. Also, in virtue of the definition of iterative actions, we know that when the agent believes he has done the iterative action, he will believe the condition is false (here, he will believe that the tree is down).

Finally, consider intending a conditional action. $(\text{INTEND}_1\ x\ [\text{IF}\ p\ \text{THEN}\ a\ \text{ELSE}\ b])$ expands into

$(\text{P-GOAL}\ x\ (\text{DONE}\ x\ [\text{BEL}\ x\ (\text{HAPPENS}\ x\ [\text{IF}\ p\ \text{THEN}\ a\ \text{ELSE}\ b])]?;$
$[\text{IF}\ p\ \text{THEN}\ a\ \text{ELSE}\ b]))$.

So, we know that eventually (unless, of course, he comes to believe the conditional is forever false), he will believe he has done the conditional in a state in which he believed he was just about to do it. By Assumption 2, an agent cannot believe he is about to do something without at least having some specific first step in mind. Thus, an agent cannot believe he is about to do a conditional involving two distinct primitive events without either believing the condition is true or believing the condition is false. So, if one intends to do a conditional action, one expects (with the usual caveats) not to be forever ignorant about the condition. This seems just right.

In summary, we have defined intending to do an action in a way that captures many reasonable properties, some of which are inherited from the commitments involved in adopting a persistent goal. However, it is often thought that one can intend to achieve states of affairs in addition to just actions. Some cases of this are discussed above. But $\text{INTEND}_1$ cannot express an agent's intending to do *something* himself to achieve a state of affairs, since the event variables are quantified outside $\text{INTEND}_1$. To allow for this case, we define another kind of intention, $\text{INTEND}_2$.

## 5.2 $\text{INTEND}_2$

One might intend to become rich, to become happy, or (perhaps controversially) to kill one's uncle,[16] without having any idea how to achieve that state of affairs, not even having an enormous disjunction of possible op-

tions. In these cases we will say the agent x is committed merely to doing something himself to bring about a world state in which (RICH x) or (HAPPY x) or (DEAD u) hold. Notice that because of the constraints that come along with adopting such a commitment, this is stronger than having only a desire or a simple goal:

*Definition 10*

$$(\text{INTEND}_2\ x\ p) \stackrel{def}{=} (\text{P-GOAL}\ x\ \exists e\ (\text{DONE}\ x\ [(\text{BEL}\ x\ \exists e'\ (\text{HAPPENS}\ x\ e';p?)) \wedge \sim(\text{GOAL}\ x\ \sim(\text{HAPPENS}\ x\ e;p?))]?;e;p?)).$$

We will explain this definition in a number of steps. First, notice that to $\text{INTEND}_2$ to bring about p, an agent is committed to doing some sequence of events e himself, after which p holds. However, as earlier, to avoid allowing an agent to intend to make p true by committing himself to doing something accidentally or unknowingly, we require the agent to think he is about to do *something* (event sequence e') bringing about p.[17] From Assumption 2 we know that even though the agent believes only that he will do *some* sequence of events achieving p, the agent will know which initial step he is about to take.

Now, it seems to us that the only way, short of truly wishful thinking, that an agent can believe he is about to do something to bring about p is if the agent in fact has a *plan* (good, bad, or ugly) for bringing it about. In general, it is quite difficult to define what a plan is, or to define what it means for an agent to have a plan.[18] The best we can do, and that is not too far off, is to say that agent must believe be is about to do something (called e' here) that will bring about p. What is left for us to specify is under what conditions this belief is *justified*, ensuring, for instance, that the agent never has such a belief when he has absolutely no idea of how to proceed.[19]

Finally, we require that prior to doing e to bring about p, the agent not have as a goal e's not bringing about p. In other words, though there may be uncertainty in the agent's mind as to which action will ultimately bring about p (for example, he may have a conditional plan), what does in fact happen had better be compatible with the agent's goals. This condition is required to handle the following example, due to Chisholm (1966) and discussed by Searle (1983). An agent intends to kill his uncle. On the way to his uncle's house, this intention causes him to become so agitated that he loses control of his car and runs over a pedestrain, who happens to be his uncle. Although the uncle is dead, we would surely say that the action that the agent did was not what was intended.

Let us cast this problem in terms of $\text{INTEND}_2$, but without the condition stating that the agent should not want e not to bring about p. Call this $\text{INTEND}_{2'}$. So, assume the following is true: $(\text{INTEND}_{2'}\ x\ (\text{DEAD}\ u))$. The agent thus has a commitment to doing some sequence of events resulting



in his uncle's death, and immediately prior to doing it he has to believe there would be some sequence (e') that he was about to do that would result in the uncle's death. However, the example satisfies these conditions, but the event he in fact does that kills his uncle may not be the one foreseen to do so. A jury requiring only the weakened $\text{INTEND}_2$ to convict for first-degree murder would find the agent to be guilty. Yet, we clearly have the intuition that the death was accidental.

Searle argues that a prior intention should cause an "intention-in-action" that presents the killing of the uncle as an "intentional object," and this causation is self-referential. To explain Searle's analysis would take us too far afield. However, we can handle this case by adding the second condition to the agent's mental state that just prior to doing the action that achieves p, he also does not want what he in fact does, e, *not* to bring about p. In the case in question (intuitively), the agent's plan is to get to his uncle's house and take it from there. Driving onto the sidewalk (and killing someone) is not one of the possible outcomes of this plan and so is ruled out by the agent's beliefs and goals. So, in swerving off the road, the agent may still have believed he was about to do something e' that would kill his uncle (and, by Proposition 17, he wanted e' to kill his uncle), but even allowing for indeterminacy in his plan, in none of his chosen worlds is his swerving off the road what kills his uncle.

Hence, our analysis predicts that the agent did not do what he intended, even though the end state was achieved, and resulted from his adopting an intention. Let us now see how this analysis stacks up against the problems and related desiderata.

### 5.3 *Meeting the Desiderata*

In this section we show how various properties of the commonsense concept of intention are captured by our analysis based on P-GOAL. In what follows we will use $\text{INTEND}_1$ or $\text{INTEND}_2$ as best fits the example. Similar results hold for analogous problems posed with the other form of intention.

We reiterate Bratman's (1983, 1987) analysis of the roles that intentions typically play in the mental life of agents:

> 1. *Intentions normally pose problems for the agent; the agent needs to determine a way to achieve them.* If the agent intends an action as described by an action expression, then the agent knows in general terms what to do. However, the action expression may have disjunctions and conditionals in it. Hence, the agent would not know at the time of forming the intention just what will be done. However, we have argued in section 5.1.1 that eventually the agent will know what actions should be taken next. In the case of nonspecific intentions,

such as (INTEND$_2$ x p), we can derive via Proposition 21 that, under the normal circumstances where the agent does not learn that p is unachievable, the agent eventually believes there is some sequence of events that he has done prior to which he believed he was about to achieve p. Hence, our analysis shows the problem that is posed by adopting a nonspecific intention, but it does not encode the solution —that the agent will form a plan specifying just what that sequence of events would be.

2. *Intentions provide a "screen of admissibility" for adopting other intentions.* If an agent has an intention to do b, and the agent (always) believes that doing a prevents the achievement of b, then the agent cannot have the intention to do a;b, or even the intention to do a before doing b. Thus, the following holds:

*Theorem 2* Screen of admissibility

$$\models \forall x\ (\text{INTEND}_1\ x\ b) \wedge \Box(\text{BEL}\ x\ [(\text{DONE}\ x\ a) \supset \Box \sim (\text{DONE}\ x\ b)]) \supset \sim(\text{INTEND}_1\ x\ a;b),$$

*where* a *and* b *are arbitrary action expressions, and their free variables have been bound outside.*

The proof is simply that there are no possible worlds in which the two intentions and the belief could all hold; in the agent's chosen worlds, if a has just been done, b will never be done. Hence, the agent cannot intend to do a before doing b. Similarly, if the agent first intends to do a and believes the above relationship between a and b, then the agent cannot also adopt the intention to do b.[20]

Notice that our agents cannot knowingly (and hence, by Proposition 18, deliberately) act against their own best interests. That is, they cannot intentionally act in a way that would make their persistent goals unachievable. Moreover, if they have adopted a time-limited intention, they cannot intend to do some other act knowing it would make achieving that time-limited intention forever false.

3. *Agents "track" the success of their attempts to achieve intentions.* In other words, agents keep their intentions after failure. Assume an agent has an intention to do a, then does something, e, thinking it will bring about the doing of a, but then comes to believe it did not. If the agent does not think that a can never be done, does the agent still have the intention to do a? Yes.

*Theorem 3*

$$\models (\text{DONE}\ x\ [(\text{INTEND}_1\ x\ a) \wedge (\text{BEL}\ x\ (\text{HAPPENS}\ x\ a))]?;e) \wedge (\text{BEL}\ x \sim(\text{DONE}\ x\ a)) \wedge \sim(\text{BEL}\ x \Box \sim(\text{DONE}\ x\ a)) \supset (\text{INTEND}_1\ x\ a).$$



The proof of this follows immediately from the definition of INTEND$_1$, which is based on P-GOAL, which states that the intention cannot be given up until it is believed to have been achieved or to be unachievable. Here, the agent believes it has not been achieved and does not believe it to be unachievable. Hence, the agent keeps the intention.

Other writers have proposed that if an agent intends to do a, then

4. *The agent believes that* a *can be done.* We do not have a modal operator for possibility. But we can state, via Proposition 17, that the agent does not believe a will never be done. This is not precisely the same as the desired property but surely is close enough for current purposes.

5. *Sometimes the agent believes he will in fact do* a. This is a consequence of Theorem 1, which states the conditions (call them C) under which $\Diamond$(DONE x a) holds, given the intention to do a. So, if the agent believes he has the intention and believes C holds, $\Diamond$(DONE x a) follows from his beliefs as well.

6. *The agent does not believe he will never do* a. This principle is embodied directly in Proposition 17, which is validated by the simple model-theoretical constraint that worlds that are consistent with one's choices are included in worlds that are consistent with one's beliefs (worlds one thinks one might be in).

7. *Agents need not intend all the expected side effects of their intentions.* Recall that in an earlier problem an agent intended to have his teeth filled. Not knowing about anaesthetics (one could assume this took place just as they were being first used in dentistry), he believed that it was always the case that if one's teeth are filled, one will feel pain. One could even say that surely the agent *chose* to undergo pain. Nonetheless, one would not like to say that he intended to undergo pain.

This problem is easily handled in our scheme: Let x be the patient. Assume p is (Filled-teeth x), and q is (In-pain x). Now, we know that the agent has surely chosen pain (by Proposition 19). Given all this, the following holds (see section 4.1.2, Case 3):

$$\not\models (\text{INTEND}_2\ x\ p) \wedge (\text{BEL}\ x\ \Box(p \supset q)) \supset (\text{INTEND}_2\ x\ q).$$

Thus, agents need not intend the expected side effects of their intentions. Contrast this, however, with a situation where the belief in the inevitability of pain is unshakeable (section 4.1.2, Case 4):

$$\models (\text{INTEND}_2\ x\ p) \wedge \Box(\text{BEL}\ x\ \Box(p \supset q)) \supset (\text{INTEND}_2\ x\ q).$$

In this case, if the patient does not experience pain, he will believe

that he has not had his tooth filled and so will persist, as with any intention.

At this point we have met the desiderata. Thus, the analysis so far has merit; but we are not finished. The definition of P-GOAL can be extended to make explicit what is only implicit in the commonsense concept of intention—the background of other justifying beliefs and intentions. Doing so will make our agents more reasonable.

## 6 *An End of Fanaticism*

As the formalism stands now, once an agent has adopted a persistent goal, he will not be deterred. For example, if agent *A* receives a request from agent *B* and decides to cooperate by adopting a persistent goal to do the requested act, *B* cannot "turn *A* off." This is clearly a defect that needs to be remedied. The remedy depends on the following definition:

*Definition 11* Persistent, relativized goal

$$(\text{P-R-GOAL x p q}) \stackrel{def}{=} (\text{GOAL x (LATER p)}) \wedge (\text{BEL x} \sim \text{p}) \wedge (\text{BEFORE [(BEL x p)} \vee (\text{BEL x} \Box \sim \text{p}) \vee (\text{BEL x} \sim \text{q})] \sim (\text{GOAL x (LATER p)})).$$

That is, a necessary condition for giving up a P-R-GOAL is that the agent x believes it is satisfied, or believes it is unachievable, or believes $\sim$q. Such propositions q form a background that justifies the agent's intentions. In many cases such propositions constitute the agent's *reasons* for adopting the intention. For example, x could adopt the persistent goal to buy an umbrella relative to his belief that it will rain. He could then consider dropping his persistent goal should he come to believe that the forecast has changed.

Our analysis supports the observation that intentions can (loosely speaking) be viewed as the contents of plans (for example, Bratman 1987; Cohen and Perrault 1979; Pollack 1986). Although we have not given a formal analysis of plans here, the commitments one undertakes with respect to an action in a plan depend on the other planned actions, as well as the pre- and postconditions brought about by those actions. If x adopts a persistent goal p relative to (GOAL x q), then necessary conditions for x's dropping his goal include his believing that he no longer has q as a goal. Thus, (P-R-GOAL x p (GOAL x q)) characterizes an agent's having a persistent *subgoal* p relative to the *supergoal* q. An agent's dropping a supergoal is now a necessary (but not sufficient) prerequisite for his dropping a subgoal.[21] Thus, with the change to relativized persistent goals, we open up the possibility of having a complex web of interdependencies among the agent's goals, intentions, and beliefs. We always had the possibility of



conditional P-GOALs. Now we have added background conditions that could lead to a revision of one's persistent goals. The definitions of intention given earlier can now be recast in terms of P-R-GOAL:

*Definition 12*

$$(\text{INTEND}_1\ \text{x a q}) \stackrel{def}{=} (\text{P-R-GOAL x [(DONE x (BEL x (HAPPENS x a))?;a)] q}).$$

*Definition 13*

$$(\text{INTEND}_2\ \text{x p q}) \stackrel{def}{=} (\text{P-R-GOAL x } \exists \text{e (DONE x [(BEL x } \exists \text{e' (HAPPENS x e';p?})) \wedge \sim(\text{GOAL x} \sim(\text{HAPPENS x e;p?}))]?;\text{e;p?}) \text{ q}).$$

With these changes, the dependencies of an agent's intentions on his beliefs, other goals, intentions, and so on, become explicit. For example, we can express an agent's intending to take an umbrella relative to believing it will rain on March 5, 1986 as

$$\exists \text{e,u (Take u e)} \wedge (\text{INTEND}_1\ \text{x e;3/5/86? } \Diamond(\text{Raining} \wedge \text{3/5/86})).$$

One can now describe agents whose primary concern is with the end result of their intentions, not so much with achieving those results themselves. An agent may first adopt a persistent goal to achieve p and then (perhaps because he does not know any other agent who will, or can, do so) subsequently decide to achieve p himself, relative to that persistent goal. So, the following is true of the agent: (P-GOAL x p) $\wedge$ ($\text{INTEND}_2$ x p (P-GOAL x p)). If someone else achieves p (and the agent comes to believe it is true), the agent must drop (P-GOAL x p) and is therefore free to drop the commitment to achieving p himself. Notice, however, that for goals that can be reachieved, the agent is *not forced* to drop the intention, as the agent may truly be committed to achieving p himself.

Matters get more interesting still when we allow the relativization conditions q to include propositions about other agents. For example, if q is (GOAL y s), then y's goal is an *interpersonal supergoal* for x. The kind of intention that is engendered by a request seems to be P-R-GOAL. Namely, the speaker tries to bring it about that

(P-R-GOAL addressee (DONE addressee a)
[GOAL speaker (DONE addressee a)]).

The addressee can get "off the hook" if he learns the speaker does not want him to do the act after all.

Notice also that given this partial analysis of requesting, a hearer who merely says "OK" and thereby accedes to a request has (made it mutually

believed that he has) adopted a commitment relative to the speaker's desires. In other words, he is committed *to* the speaker to do the requested action. This helps to explain how social commitments can arise out of communication. However, this is not the place to analyze speech acts (but see chapter 12 in this volume).

Finally, interlocking commitments are obtained when two agents are in the following states: (P-R-GOAL x p (GOAL y p)), and (P-R-GOAL y p (GOAL x p)). Each agent will keep his intention at least as long as the other keeps it. For example, each might have the intention to lift a table. But each would not bother to try unless the other also had the same intention.[22]

In summary, persistent relativized goals provide a useful analysis of intention and extend the commonsense concept by making explicit the conditions under which an agent will revise his intentions.

## 7 *Conclusion*

This paper establishes basic principles governing the rational balance among an agent's beliefs, actions, and intentions. Such principles provides specifications for artificial agents and approximate a theory of human action (as philosophers use the term). By making explicit the conditions under which an agent can drop his goals—that is, by specifying how the agent is *committed* to his goals—the formalism captures a number of important properties of intention. Specifically, the formalism provides analyses for Bratman's (1983, 1987) three characteristic functional roles played by intentions and shows how agents can avoid intending all the foreseen side effects of what they actually intend. Finally, the analysis shows how intentions can be adopted relative to a background of relevant beliefs and other intentions or goals. By relativizing one agent's intentions in terms of beliefs about another agent's intentions (or beliefs), we derive a preliminary account of interpersonal commitments.

The utility of the theory for describing people or artificial agents will depend on the fidelity of the assumptions. It does not seen unreasonable to require that a robot not procrastinate forever. Moreover, we surely would want a robot to be persistent in pursuing its goals, but not fanatically so. Furthermore, we would want a robot to drop goals given to it by other agents when it determines the goals need not be achieved. So, as a coarse description of an artificial agent, the theory seems workable.

The theory is not only useful for describing single agents in dynamic multiagent worlds, it is also useful for describing their interactions, especially via the use of communicative acts. In a companion paper (see chapter 12 in this volume) we present a theory of speech acts that builds on the foundations laid here.



Much work remains. The action theory only allows for possible worlds consisting of single courses of events. Further developments should include basing the analysis on partial worlds/situations (Barwise and Perry 1983) and on temporal logics that allow for simultaneous action (Allen 1984; Georgeff 1987; Lansky 1987). Finally, the theory would be strengthened by the use of default and nonmonotonic reasoning.

### *Notes*

1. We thank Nils Nilsson for this apt phrase.
2. Exceptions include the work of Moore (1980), who analyzed the relationship of knowledge to action, and that of Appelt (1981) and Konolige (1980, 1985). However, none of these works addressed the issue of goals and intention.
3. Rosenschein (1981) discusses some of the difficulties of hierarchical planners and presents a formal theory of plans in terms of dynamic logic.
4. Or her. We use the masculine version here throughout.
5. The rationale for this property was discussed above.
6. Many theories of intention are committed to the undesirable view that expected side effects to one's intentions are intended as well.
7. Such desires are ones that speech act theorists claim to be conveyed by illocutionary acts such as requests.
8. This is not a *social* commitment. It remains to be seen whether the latter can be built out of the former.
9. This is unlike the integration of similar operators by Moore (1980), who analyzes how an agent's knowledge affects and is affected by his actions. That research meshed a possible-worlds model of knowledge with a situation-calculus-style, branching-time model of action (McCarthy and Hayes 1969). Our earlier work (Cohen and Levesque 1985) used a similar branching-time/dynamic logic model. However, the model's inability to express beliefs about what was in fact about to happen in the future led to many difficulties.
10. Without this choice, it is far from clear that any kind of coherent action would be possible.
11. One also needs to show that there is at least one model that satisfies these assumptions. This is straightforward.
12. For an exploration of the issues involved in explicit versus implicit belief, see Levesque 1984.
13. However, we have assumed immortal agents.
14. For example, someone may be committed to your knowing q but not to achieving q itself.
15. Notice that e could be the last step of a.
16. We are not trying to be morbid here, just setting up a classic example.
17. The definition does not use e instead of e' because that would quantify e into the agent's beliefs, requiring that he (eventually) have picked out a precise sequence of events that he thinks will bring about p. If we wanted to do that, we could use $\text{INTEND}_1$.
18. See Pollack 1986 for a discussion of these issues.
19. One possibility is to make sure this belief *only* arises by existential generalization from a belief involving a particular action description (that is, the plan) achieving p. However, one cannot express this constraint in our logic since one cannot quantify over action expressions.

20. Notice that the theorem does not require quantification over primitive acts but allows a and b to be arbitrary action expressions.
21. Also, notice that (P-GOAL x p) is now subsumed by (P-R-GOAL x p $\sim$p).
22. Ultimately, one can envision circular interlinkages in which one agent adopts a persistent goal provided another agent has adopted it relative to the first agent's having adopted it relative to the second agent's having adopted it, and so on. For an analysis of circular propositions that might make such concepts expressible, see Barwise and Etchemendy 1987.

## *References*


Allen, J. F. (1979). A plan-based approach to speech act recognition. Technical Report 121, Department of Computer Science, University of Toronto, Toronto, Ont.

Allen, J. F. (1984). Towards a general theory of action and time. *Artificial Intelligence* 23, 123–154.

Allen, J. F., and C. R. Perrault (1980). Analyzing intention in utterances. *Artificial Intelligence* 15, 143–178.

Appelt, D. (1981). Planning natural language utterances to satisfy multiple goals. Doctoral dissertation, Stanford University, Stanford, CA.

Barwise, J., and J. Etchemendy (1987). *The liar: An essay on truth and circularity*. New York: Oxford University Press.

Barwise, J., and J. Perry (1983). *Situations and attitudes*. Cambridge, MA: MIT Press.

Bratman, M. (1983). Castañeda's theory of thought and action. In J. Tomberlin, ed., *Agent, language, and the structure of the world: Essays presented to Hector-Neri Castañeda with his replies*. Indianapolis, IN: Hackett.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review* 93, 375–405.

Bratman, M. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Castañeda, H.-N. (1975). *Thinking and doing*. Dordrecht, Holland: D. Reidel.

Chisholm, R. M. (1966). Freedom and action. In K. Lehrer, ed., *Freedom and determinism*. New York: Random House.

Cohen, P. R., and H. J. Levesque (1980). Speech acts and the recognition of shared plans. In *Proceedings of the Third Biennial Conference*, Canadian Society for Computational Studies of Intelligence, Victoria, B.C.

Cohen, P. R., and H. J. Levesque (1985). Speech acts and rationality. In *Proceedings of the Twenty-third Annual Meeting*, Association for Computational Linguistics, Chicago, IL.

Cohen, P. R., and C. R. Perrault (1979). Elements of a plan-based theory of speech acts. *Cognitive Science* 3, 177–212. Reprinted in B. Webber and N. Nilsson, eds. (1981). *Readings in artificial intelligence*. Los Altos, CA: Morgan Kaufmann. Also in B. G. Grosz, K. Sparck Jones, and B. Webber, eds. (1986). *Readings in natural language processing*. Los Altos, CA: Morgan Kaufmann.

Fagin, R., and J. Y. Halpern (1985). Belief, awareness, and limited reasoning: Preliminary report. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, CA.

Fikes, R., and N. J. Nilsson (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2, 189–208.

Georgeff, M. P. (1983). Communication and interaction in multi-agent planning. In *Proceedings of the National Conference*, American Association for Artificial Intelligence, Washington, DC.

Georgeff, M. P. (1987). Actions, processes, and causality. In *Reasoning about actions and plans: Proceedings of the 1986 workshop*. Los Altos, CA: Morgan Kaufmann.





Georgeff, M. P., and A. L. Lansky (in preparation). A BDI semantics for the procedural reasoning system. Technical Note, Artificial Intelligence Center, SRI International, Menlo Park, CA.

Halpern, J. Y., and Y. O. Moses (1985). A guide to the modal logics of knowledge and belief. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, CA.

Harel, D. (1979). *First-order dynamic logic*. New York: Springer-Verlag.

Harman, G. (1986). *Change in view*. Cambridge, MA: MIT Press.

Konolige, K. (1980). A first-order formalization of knowledge and action for a multiagent planning system. Technical Note 232, Artificial Intelligence Center, SRI International, Menlo Park, CA. (Appears in *Machine Intelligence* 10.)

Konolige, K. (1985). Experimental robot psychology. Technical Note 363, Artificial Intelligence Center, SRI International, Menlo Park, CA.

Konolige, K., and N. J. Nilsson (1980). Multiple-agent planning systems. In *Proceedings of the National Conference*, American Association for Artificial Intelligence, Stanford, CA.

Lamport, L. (1980). "Sometimes" is sometimes better than "not never." In *Proceedings of the Seventh Annual ACM Symposium on Principles of Programming Languages*, Association for Computing Machinery.

Lansky, A. L. (1985). Behavioral specification and planning for multiagent domains. Technical Note 360. Artificial Intelligence Center, SRI International, Menlo Park, CA.

Lansky, A. L. (1987). A representation of parallel activity based on events, structure, and causality. In *Reasoning about actions and plans: Proceedings of the 1986 workshop*. Los Altos, CA: Morgan Kaufmann.

Levesque, H. J. (1984). A logic of implicit and explicit belief. In *Proceedings of the National Conference*, American Association for Artificial Intelligence, Austin, TX.

McCarthy, J., and P. J. Hayes (1969). Some philosophical problems from the standpoint of artificial intelligence. In *Machine intelligence 4*. New York: American Elsevier.

Moore, R. C. (1980). Reasoning about knowledge and action. Technical Note 191, Artificial Intelligence Center, SRI International, Menlo Park, CA.

Pollack, M. E. (1986). Inferring domain plans in question-answering. Doctoral dissertation, Department of Computer Science, University of Pennsylvania, Philadelphia, PA.

Pratt, V. R. (1978). Six lectures on dynamic logic. Technical Report MIT/LCS/TM-117, Laboratory for Computer Science, MIT, Cambridge, MA.

Rosenschein, S. J. (1981). Plan synthesis: A logical perspective. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Vancouver, B.C.

Rosenschein, J. S. (1986). Rational interaction: Cooperation among intelligent agents. Doctoral dissertation, Department of Computer Science, Stanford University, Stanford, CA.

Rosenschein, J. S., and M. R. Genesereth (1984). Communication and cooperation. Technical Report 84–5, Heuristic Programming Project, Department of Computer Science, Stanford University, Stanford, CA.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. New York: Cambridge University Press.