Intention Is Choice with Commitment*

Philip R. Cohen
Artificial Intelligence Center and Center for the Study of Language and Information, SRI International, Menlo Park, CA 94025, USA

Hector J. Levesque**
Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A4

ABSTRACT

This paper explores principles governing the rational balance among an agent's beliefs, goals, actions, and intentions. Such principles provide 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, i.e., 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 three characteristic functional roles played by intentions [7, 9], *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.*

1. Introduction

Some time in the not-so-distant future, you are having trouble with your new

*This research was made possible by a gift from the Systems Development Foundation, by support from the Natural Sciences and Engineering Research Council of Canada, and 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, the United States Government, or the Canadian Government. Earlier versions of this paper have appeared in *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop at Timberline Lodge* (Morgan Kaufmann, Los Altos, CA), in *Proceedings AAAI-87*, Seattle, WA, and in *Intentions in Communication*, edited by P.R. Cohen, J. Morgan and M.E. Pollack (MIT Press, Cambridge, MA).

**Fellow of the Canadian Institute for Advanced Research.

Artificial Intelligence **42** (1990) 213–261
0004-3702/90/$3.50

household robot.[1] You say "Willie, bring me a beer." The robot replies "OK, boss." Twenty minutes later, you screech "Willie, why didn't you bring that beer?" It answers "Well, I intended to get you the beer, but I decided to do something else." Miffed, you send the wise guy back to the manufacturer, complaining about a lack of commitment. After retrofitting, Willie is returned, marked "Model C: The Committed Assistant." Again, you ask Willie to bring a beer. Again, it accedes, replying "Sure thing." Then you ask: "What kind did you buy?" It answers: "Genessee." You say "Never mind." One minute later, Willie trundles over with a Genessee in its gripper. This time, you angrily return Willie for overcommitment. After still more tinkering, the manufacturer sends Willie back, promising no more problems with its commitments. So, being a somewhat trusting consumer, you accept the rascal back into your household, but as a test, you ask it to bring you your last beer. Willie again accedes, saying "Yes, Sir." (Its attitude problem seems to have been fixed.) The robot gets the beer and starts towards you. As it approaches, it lifts its arm, wheels around, deliberately smashes the bottle, and trundles off. Back at the plant, when interrogated by customer service as to why it had abandoned its commitments, the robot replies that according to its specifications, it kept its commitments as long as required—commitments must be dropped when fulfilled or impossible to achieve. By smashing the last bottle, the commitment became unachievable.

Despite the impeccable logic, and the correct implementation, Willie is dismantled.

1.1. Rational balance

This paper is concerned with specifying the "rational balance"[2] needed among the beliefs, goals, plans, intentions, commitments, and actions of autonomous agents. Our specific objective is to explore the relationship that intention plays in maintaining this balance. For example, the following are desirable properties of intention: An autonomous agent should act on its intentions, not in spite of them; adopt intentions it believes are feasible and forego those believed to be infeasible; keep (or commit to) intentions, but not forever; discharge those intentions believed to have been satisfied; alter intentions when relevant beliefs change; and adopt subsidiary intentions during plan formation. From this list one can see that an agent's adopting an intention has many effects on its mental state. To specify what it means for an agent to have an intention, one needs to describe how that intention affects the agent's web of beliefs, commitments to future actions, and other interdependent intentions.

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

[1] This problematic robot is very loosely based on Willie, the robot in Philip K. Dick's novel, *The Galactic Pot-Healer*.

[2] We thank Nils Nilsson for this apt phrase.

rational balance is to examine various commonsense relationships among human beliefs, intentions, and commitments that seem to justify our attribution of the term "rational." However, our goals in this study extend to the characterization of interaction and communication. We propose a logic suitable both for describing and reasoning about agent's mental states as well as agents' abilities to affect the mental states of others. Not only will a theorist have to reason about the kinds of interactions agents can have, in order to communicate, agents themselves need to reason about the beliefs, intentions, and commitments of other agents. Although we do not consider communication here, our analysis serves as a foundation for a theory of speech acts [13, 14], and applies more generally to situations in which communication may take place in an artificial 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.

1.2. Artificial intelligence research on planning systems

AI research has concentrated on algorithms for finding plans to achieve given goals, on monitoring plan execution, and on replanning [18]. Recently, planning in dynamic, multiagent 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 [1, 3, 4, 12, 13, 15, 19, 21, 29, 43, 44]. 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.[3] 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 way of operationalizing the concept of intention has a number of difficulties. First, although there surely *is* a strong relationship between plans and intentions [41], agents may form plans that they never "adopt," and thus the notion of a plan lacks the characteristic commitment to action inherent in our commonsense understanding of intention. Second, even if we accept the claim that a planning system's intentions are the contents of its plans, what constitutes a plan for most planning systems is itself often a murky

[3] Exceptions include the work of Moore [36] who analyzed the relationship of knowledge to action, and that of Appelt [4], Haas [23], Konolige [27, 28] and Morgenstern [37]. However, none of these works address the issue of goals and intention.

topic.[4] Thus, saying that the system's intentions are the contents of its plans lacks needed precision. Finally, 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. Communication involves one's ability to reason about the intentions of others [22, 39]. With only an operational definition rather than a declarative characterization of rational balance, it becomes quite difficult to engage in such reasoning. 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.

1.3. 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 shall explore their territory here, but 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 [8, 9, 47]. The former guide agents' planning and constrain their adoption of other intentions [9], whereas the latter function *causally* in producing behavior [47]. 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 analysis has examined the relationship between an agent's doing something intentionally and that agent's having a present-directed intention. Recently, Bratman [8] 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 [10] treats the content of an intention as a "practition," similar to an action description (in computer science terms). 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

[4] Rosenschein [45] identifies some of the semantic weaknesses in the AI literature's treatment of hierarchically specified plans, and presents a formal theory of plans in terms of dynamic logic.

also been argued that needed connections between propositions and practitions may not be derivable [7].

Searle [47] claims that the content of an intention is a causally self-referential representation of its conditions of satisfaction (and see also [26]). 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 characterize 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 sections, we explore what it is that theories of intention should handle.

1.4. Desiderata for a theory of intention

Bratman [9] 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[5] 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 shall 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 in California intends to fly to New York on a certain date, he should be motivated to find his way to

[5] Or her: we use the masculine version here throughout.

New York. That is, the agent should form a plan of action to go to New York, and then (all else being equal) to do something in order to get there. If the agent takes no actions to enable him to do so, then the intention did not affect 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.[6]

(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.*[7]

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.

[6] The rationale for this property was discussed above.

[7] Many theories of intention are committed to the undesirable view that expected side-effects to one's intentions are intended as well.

We will develop a theory in which expected side-effects are *chosen*, but not intended. These properties are our primary desiderata for a treatment of intention. However, motivated by AI research, we add one other, as described below:

1.5. The "Little Nell" problem: Not giving up too soon

McDermott [35] points out the following difficulty with a naively designed planning system:

> Say a problem solver is confronted with the classic situation of a heroine, called Nell, having been tied to the tracks while a train approaches. The problem solver, called Dudley, knows that "If Nell is going to be mashed, I must remove her from the tracks." (He probably knows a more general rule, but let that pass.) When Dudley deduces that he must do something, he looks for, and eventually executes, a plan for doing it. This will involve finding out where Nell is, and making a navigation plan to get to her location. Assume that he knows where she is, and he is not too far away; then the fact that the plan will be carried out will be added to Dudley's world model. Dudley must have some kind of data-base-consistency maintainer (Doyle, 1979) to make sure that the plan is deleted if it is no longer necessary. Unfortunately, as soon as an apparently successful plan is added to the world model, the consistency maintainer will notice that "Nell is going to be mashed" is no longer true. But that removes any justification for the plan, so it goes too. But that means "Nell is going to be mashed" is no longer contradictory, so it comes back in. And so forth. (p.102)

The agent continually plans to save Nell, and abandons its plan because it believes it will be successful. McDermott attributes the problem to the inability of various planning systems to express "Nell is going to be mashed *unless* I save her," and to reason about the concept of prevention. Haas [23] blames the problem on a failure to distinguish between actual and possible events. The planner should be trying to save Nell based on a belief that it is possible that she will be mashed, rather than on the belief that she in fact will be mashed. Although reasoning about prevention, expressing "unless," and distinguishing between possible and actual events are important aspects of the original formulation of the problem, the essence of the Little Nell problem is the more general problem of an agent's giving up an intention too soon. We shall show how to avoid it.

As should be clear from the previous discussion, much rides on an analysis of intention and commitment. In the next section, we indicate how these concepts can be approximated.

1.6. Intention 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.[8] 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 (i.e., 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 fail, 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.[9] 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 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 theory of 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

[8] Such desires are ones that speech act theorists claim to be conveyed by illocutionary acts such as requests.

[9] This is not a *social* commitment. It remains to be seen if the latter can be built out of the former.

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 [14]), 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 analyses 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 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 carefully worked out 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 shall 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 [30, 31]).[10] By adding GOAL, we can model an agent's intentions.

Intuitively, a model for these operators includes courses of events, which consist of sequences of primitive events, that characterize what has happened and will happen in each possible world.[11] Possible worlds can also be related to one another via accessibility relations that partake in the semantics of BEL and GOAL. 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.

As a general strategy, the formalism will be too strong. First, we have the usual consequential closure problems that plague possible-worlds models for belief. These, however, will be accepted for the time being, and we welcome

[10] This is unlike the integration of similar operators by Moore [36], 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 [34]. Our earlier work [13] used a similar branching-time dynamic-logic model. However, the model's inability to support beliefs about what was in fact about to happen in the future led to many difficulties.

[11] For this paper, the only events that will be considered are those performed by an agent. These events may be thought of as event types, in that they do not specify the time of occurrence, but do include all the other arguments. Thus John's hitting Mary would be such an event type.

attempts to develop finer-grained semantics (e.g., [6, 16]). 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 non-monotonic reasoning could smooth over the exceptions, but we will not attempt to specify such reasoning here (but see [38]). 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 dropped too soon.

3.1. Syntax

For simplicity, we adopt a logic with no singular terms, using instead predicates and existential quantifiers. However, for readability, we will often use constants. The interested reader can expand these out into the full predicative form if desired.

⟨**Action-var**⟩ ::= $a, b, a_1, a_2, \ldots, b_1, b_2, \ldots, e, e_1, e_2, \ldots$.

⟨**Agent-var**⟩ ::= $x, y, x_1, x_2, \ldots, y_1, y_2, \ldots$.

⟨**Regular-var**⟩ ::= $i, j, i_1, i_2, \ldots, j_1, j_2, \ldots$.

⟨**Variable**⟩ ::= ⟨Agent-var⟩ | ⟨Action-var⟩ | ⟨Regular-var⟩ .

⟨**Pred**⟩ ::= (⟨Pred-symbol⟩ ⟨Variable⟩$_1, \ldots,$ ⟨Variable⟩$_n$) .

⟨**Wff**⟩ ::= ⟨Pred⟩ | ¬⟨Wff⟩ | ⟨Wff⟩ ∨ ⟨Wff⟩ | ∃⟨Variable⟩ ⟨Wff⟩ |
one of the following:

⟨Variable⟩ = ⟨Variable⟩,

(HAPPENS ⟨Action-expression⟩):
⟨Action-expression⟩ happens next,

(DONE ⟨Action-expression⟩):
⟨Action-expression⟩ has *just* happened,

(AGT ⟨Agent-var⟩ ⟨Action-var⟩):
⟨Agent-var⟩ is the *only* agent of ⟨Action-var⟩,

(BEL ⟨Agent-var⟩ ⟨Wff⟩):
⟨Wff⟩ follows from ⟨Agent-var⟩'s beliefs,

(GOAL ⟨Agent-var⟩ ⟨Wff⟩):
⟨Wff⟩ follows from ⟨Agent-var⟩'s goals,

⟨Time-proposition⟩,
⟨Action-var⟩ ≤ ⟨Action-var⟩.

⟨**Time-proposition**⟩ ::= ⟨Numeral⟩ (see below)

⟨**Action-expression**⟩ ::= ⟨Action-var⟩ | one of the following:
⟨Action-expression⟩; ⟨Action-expression⟩: sequential action,
⟨Action-expression⟩ | ⟨Action-expression⟩:
nondeterministic choice action,
⟨Wff⟩?: test action,
⟨Action-expression⟩*: iterative action.

Time propositions are currently just numerals. However, for ease of exposition, we shall write them as if they were time-date expressions such as 2:30PM/3/6/85. These will be true or false in a course of events at a given index iff the index is the same as that denoted by the time proposition (i.e., 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) ∧ 2.30PM/3/6/85. Thus, if the above conjunction were a goal, both conjuncts would have to be true simultaneously.

3.2. Semantics

We shall adapt the usual possible-worlds model for belief to deal with goals. Assume there is a set of possible worlds T, each one consisting of a sequence (or course) or events, temporally extended infinitely in past and future. Each possible world characterizes possible ways the world could have been, and could be. Thus, each world specifies what happens in the future. Agents usually do not know precisely which world they are in. Instead, some of the worlds in T are consistent with the agents beliefs, and some with his goals, where the consistency is specified in the usual way, by means of an accessibility relation on tuples of worlds, agents, and an index, n, into the course of events defining the world (from which one can compute a time point, if need be).

To consider what an agent believes (or has as a goal), one needs to supply a world and an index into the course of events defining that world. As the world evolves, agents' beliefs and goals change. When an agent does an action in some world, he does not bring about a new world, though he can alter the facts of that world at that time. Instead, after an event has happened, we shall say the world is in a new "state" in which new facts hold and the set of accessible worlds has been altered. That is, the agent changes the way he *thinks* the world could be and/or chooses the world to be.

3.2.1. *Model theory*

A model M is a structure $\langle \Theta, P, E, Agt, T, B, G, \Phi \rangle$, where Θ is a set, P is a set of people, E is a set of primitive event types, $Agt \in [E \rightarrow P]$ specifies the agent of an event, $T \subseteq [\mathbb{Z} \rightarrow E]$ is a set of possible courses of events (or worlds) specified as a function from the integers to elements of E, $B \subseteq T \times P \times \mathbb{Z} \times T$ is the belief accessibility relation, $G \subseteq T \times P \times \mathbb{Z} \times T$ is the goal accessibility relation, and Φ interprets predicates. Formulas will be evaluated with respect to some possible course of events, hereafter some *possible world*, and an "index" into that possible world, that is, at a particular point in the course of events.[12]

3.2.2. *Definitions*

(1) $D = \Theta \cup P \cup E^*$, specifying the domain of quantification. That is, one can quantify over things, people, and sequences of (types of) primitive events. Given this, $\Phi \subseteq [Pred^k \times T \times \mathbb{Z} \times D^k]$.
(2) $AGT \subseteq E^* \times P$, where $x \in AGT[e_1, \ldots, e_n]$ iff there is an i such that $x = Agt(e_i)$. That is, AGT specifies the partial agents of a sequence of events.

3.2.3. *Satisfaction*

Assume M is a model, σ a sequence of events, n an integer, v a set of bindings of variables to objects in D, and if $v \in [Vars \rightarrow D]$, then v^{x}_d is that function which yields d for x and is the same as v everywhere else. We now specify what it means for M,σ,v,n to *satisfy* a wff α, which we write as $M,\sigma,v,n \models \alpha$. Because of formulas involving actions, this definition depends on what it means for an expression a to *occur* between index points n and m. This, we write as $M,\sigma,v,n[\![\mathsf{a}]\!]m$, and is itself defined in terms of satisfaction. The definitions are as follows:

(1) $M,\sigma,v,n \models P(\mathsf{x}_1, \ldots, \mathsf{x}_k)$ iff $\langle v(\mathsf{x}_1) \ldots v(\mathsf{x}_k) \rangle \in \Phi[P, \sigma, n]$. Notice that the interpretation of predicates depends on the world σ and the event index n.
(2) $M,\sigma,v,n \models \neg\alpha$ iff $M,\sigma,v,n \not\models \alpha$.
(3) $M,\sigma,v,n \models (\alpha \vee \beta)$ iff $M,\sigma,v,n \models \alpha$ or $M,\sigma,v,n \models \beta$.
(4) $M,\sigma,v,n \models \exists \mathsf{x}\alpha$ iff $M,\sigma,v^{\mathsf{x}}_d,n \models \alpha$ for some d in D.
(5) $M,\sigma,v,n \models (\mathsf{x}_1 = \mathsf{x}_2)$ iff $v(\mathsf{x}_1) = v(\mathsf{x}_2)$.
(6) $M,\sigma,v,n \models \langle$Time-proposition$\rangle$ iff $v(\langle$Time-proposition$\rangle) = n$.

Next, we treat events and actions, describing what it means for an action to be about to occur, and to have just occurred:

[12] For those readers accustomed to specifying possible worlds as real-numbered times and events as denoting intervals over them (e.g., [2]), we remark that we shall not be concerned in this paper about parallel execution of events over the same time interval. Hence, we model possible worlds as courses (i.e., sequences) of events.

(1) $M,\sigma,v,n \models (\mathsf{e}_1 \leq \mathsf{e}_2)$ iff $v(\mathsf{e}_1)$ is an initial subsequence of $v(\mathsf{e}_2)$.
(2) $M,\sigma,v,n \models (\mathsf{AGT\ x\ e})$ iff $AGT[v(\mathsf{e})] = \{v(\mathsf{x})\}$. AGT thus specifies the *only* agent e.
(3) $M,\sigma,v,n \models (\mathsf{HAPPENS\ a})$ iff $\exists m,\ m \geqslant n$, such that $M,\sigma,v,n[\![\mathsf{a}]\!]m$. That is, a describes a sequence of events that happens "next" (after n).
(4) $M,\sigma,v,n \models (\mathsf{DONE\ a})$ iff $\exists m,\ m \leqslant n$, such that $M,\sigma,v,m[\![\mathsf{a}]\!]n$. That is, a describes a sequence of events that *just* happened (before n).

Notice that the semantics of DONE and HAPPENS depends on the relation $[\![\ ]\!]$, which describes when an action occurs between two points in time. Next, we provide a semantics for statements about beliefs and goals:

(1) $M,\sigma,v,n \models (\mathsf{BEL\ x}\ \alpha)$ iff for all σ^* such that $\langle \sigma, n \rangle B[v(\mathsf{x})]\sigma^*$, $M,\sigma^*,v,n \models \alpha$. That is, α *follows from* the agents beliefs iff α is true in all possible worlds accessible via B, at index n.
(2) $M,\sigma,v,n \models (\mathsf{GOAL\ x}\ \alpha)$ iff for all σ^* such that $\langle \sigma, n \rangle G[v(\mathsf{x})]\sigma^*$, $M,\sigma^*,v,n \models \alpha$. Similarly, α *follows from* the agent's goals iff α is true in all possible worlds accessible via G, at index n.

Turning now to the occurrence of actions, we have the following definition of $[\![\ ]\!]$, describing when it can be said that a complex action "occurs" between two time points:

(1) $M,\sigma,v,n[\![\mathsf{e}]\!]n + m$ (where e is an event variable) iff $v(\mathsf{e}) = e_1 e_2 \ldots e_m$ and $\sigma(n + i) = e_i$, $1 \leqslant i \leqslant m$. Intuitively, e denotes some sequence of events of length m which appears next after n in the world σ.
(2) $M,\sigma,v,n[\![\mathsf{a|b}]\!]m$ iff $M,\sigma,v,n[\![\mathsf{a}]\!]m$ or $M,\sigma,v,n[\![\mathsf{b}]\!]m$. Either the action described by a or that described by b occurs within the interval.
(3) $M,\sigma,v,n[\![\mathsf{a;b}]\!]m$ iff $\exists k,\ n \leqslant k \leqslant m$, such that $M,\sigma,v,n[\![\mathsf{a}]\!]k$ and $M,\sigma,v,k[\![\mathsf{b}]\!]m$. The action described by a and then that described by b occurs.
(4) $M,\sigma,v,n[\![\alpha?]\!]n$ iff $M,\sigma,v,n \models \alpha$. The test action, $\alpha?$, involves no events at all, but occurs if α holds, or "blocks" (fails), when α is false. Thus, to say a test action of wff α occurred at some time point n is merely a way of constraining the course of events to be one in which α holds at n. Notice that here $[\![\ ]\!]$ is mutually recursive with $\models$.
(5) $M,\sigma,v,n[\![\mathsf{a}^*]\!]m$ iff $\exists n_1, \ldots, n_k$ where $n_1 = n$ and $n_k = m$ and for every i such that $1 \leqslant i \leqslant m$, $M,\sigma,v,n_i[\![\mathsf{a}]\!]n_{i+1}$. The iterative action a^* occurs between n and m provided only a sequence of what is described by a occurs within the interval.

A wff α is *satisfiable* if there is at least one model M, world σ, index n, and value assignment v such that $M,\sigma,v,n \models \alpha$. A wff α is *valid*, iff for every model M, world σ, event index n, and assignment of variables v, $M,\sigma,v,n \models \alpha$. To simplify the exposition, we may express the fact that a wff α is valid by $\models \alpha$.

3.2.4. *Abbreviations*

It will be convenient to adopt the following:

- *Empty sequence*: $\mathsf{nil} \stackrel{\text{def}}{=} (\forall \mathsf{x}\,(\mathsf{x} = \mathsf{x}))?$ and $\mathsf{a} = \mathsf{NIL} \stackrel{\text{def}}{=} \forall \mathsf{b}\,(\mathsf{a} \leqslant \mathsf{b})$.
  As a test action, NIL always succeeds; as an event sequence, it is a subsequence of every other one.
- *Conditional action*: $[\mathsf{IF}\ \alpha\ \mathsf{THEN\ a\ ELSE\ b}] \stackrel{\text{def}}{=} \alpha?;\mathsf{a} \mid \neg\alpha?;\mathsf{b}$.
  That is, as in dynamic logic, an if-then-else action is a disjunctive action of doing action a at a time at which α is true or doing action b at a time at which α is false. Note that the semantics of a conditional action does not require that the condition be believed by someone to be true. However as will be discussed later, when agents execute conditionals with disjoint branches, they will have to believe the condition is true (or believe it is false).
- *While-loops*: $[\mathsf{WHILE}\ \alpha\ \mathsf{DO\ a}] \stackrel{\text{def}}{=} (\alpha?;\mathsf{a})^*;\neg\alpha?$
  While-loops are a sequence of doing action a zero or more times, prior to each of which α is true. After the iterated action stops, α is false.
- *Eventually*: $\Diamond\alpha \stackrel{\text{def}}{=} \exists x\,(\mathsf{HAPPENS}\ x;\alpha?)$.
  In other words, $\Diamond\alpha$ is true (in a given possible world) if there is some sequence of events after which α will hold, that is, if α is true at some point in the future.
- *Always*: $\Box\alpha \stackrel{\text{def}}{=} \neg\Diamond\neg\alpha$.
  $\Box\alpha$ means that α is henceforth true in the course of events. A useful application of $\Box$ is $\Box(\mathsf{p} \supset \mathsf{q})$, in which no matter what happens, p still implies q. We can now distinguish between $\mathsf{p} \supset \mathsf{q}$'s being logically valid, its being true in all courses of events, and its merely being true after some event happens.

3.2.5. *Constraints on the model*

(1) *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 [24]. 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.

(2) *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 for himself to go 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 nonlogical axioms that constrain the models that we consider.

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 which 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 $\neg$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 $\neg$(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 say 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 (e.g., [25, 36, 42]), including the following:

Proposition 3.1. *Properties of complex acts*:

$$\models (\text{HAPPENS a;b}) \equiv (\text{HAPPENS a;(HAPPENS b)?}) ,$$
$$\models (\text{HAPPENS a|b}) \equiv (\text{HAPPENS a}) \vee (\text{HAPPENS b}) ,$$
$$\models (\text{HAPPENS p?;q?}) \equiv p \wedge q ,$$
$$\models (\text{HAPPENS a*;b}) \equiv (\text{HAPPENS b|a;a*;b}) .$$

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 test 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 the a*;b again.

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

Proposition 3.2. $\models (\text{HAPPENS a}) \equiv (\text{HAPPENS a;(DONE a)?})$

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

Proposition 3.3. $\models (\text{DONE a}) \equiv (\text{DONE (HAPPENS a)?;a})$

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 shall 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 3.4. $\models p \equiv (\text{DONE p?})$

That is, the test action filters out courses of events in which the proposition tested is false. The truth of Proposition 3.4 follows immediately from the definition of "?".

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

Definition 3.5. $(\text{DONE x a}) \stackrel{\text{def}}{=} (\text{DONE a}) \wedge (\text{AGT x a})$

Definition 3.6. $(\text{HAPPENS x a}) \stackrel{\text{def}}{=} (\text{HAPPENS a}) \wedge (\text{AGT x a})$.

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 $\exists e\,(\text{DONE p?;e})$. $\Diamond$ 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 $\Diamond \neg p$ are jointly satisfiable. Since $\Diamond p$ starts "now," the following property is also true,

Proposition 3.7. $\models p \supset \Diamond p$.

The following are also trivial consequences:

Proposition 3.8. $\models \Diamond(p \vee q) \wedge \Box \neg q \supset \Diamond p$,

Proposition 3.9. $\models \Box(p \supset q) \wedge \Diamond p \supset \Diamond q$.

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

Definition 3.10. $(\text{LATER p}) \stackrel{\text{def}}{=} \neg p \wedge \Diamond p$.

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

Proposition 3.11. $\models \neg(\text{LATER } \Diamond p)$.

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 3.12.

$$(\text{BEFORE p q}) \stackrel{\text{def}}{=} \forall c\,(\text{HAPPENS c;q?}) \supset \exists a\,(a \leqslant c) \wedge (\text{HAPPENS a;p?})\,.$$

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 3.13. $\models \Diamond q \wedge (\mathsf{BEFORE}\ p\ q) \supset \Diamond p$.

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 3.14. $\models \neg p \supset (\mathsf{BEFORE}\ (\exists e\ (\mathsf{DONE}\ \neg p?;e;p?))\ p)$.

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.[13] 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 be 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.

3.5. Belief

For simplicity, we assume the usual Hintikka-style axiom schemata for BEL [24] (corresponding to a "Weak S5" modal logic).

Proposition 3.15. *Belief axioms*:

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

And, we have the usual "necessitation" rule:

Proposition 3.16. *If* $\models \alpha$, *then* $\models (\mathsf{BEL}\ x\ \alpha)$.

If α is a theorem (i.e., 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:

[13] For an exploration of the issues involved in explicit versus implicit belief, see [33].

Proposition 3.17. *If* $\models \alpha$, *then* $\models (\mathsf{BEL}\ x\ \Box\alpha)$.

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

Definition 3.18. $(\mathsf{KNOW}\ x\ p) \stackrel{\text{def}}{=} p \wedge (\mathsf{BEL}\ x\ p)$.

Of course, this characterization of knowledge has many known difficulties, but 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 3.19. $(\mathsf{COMPETENT}\ x\ p) \stackrel{\text{def}}{=} (\mathsf{BEL}\ x\ p) \supset (\mathsf{KNOW}\ x\ p)$.

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 3.20. $\models \forall x,e\ (\mathsf{AGT}\ x\ e) \supset [(\mathsf{DONE}\ e) \equiv (\mathsf{BEL}\ x\ (\mathsf{DONE}\ e))]$.

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*. In Section 5.1, we will make additional assumptions about the actions an agent is about to perform.

3.6. 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 3.21. *Consistency*: $\models \forall x\ (\mathsf{GOAL}\ x\ p) \supset \neg(\mathsf{GOAL}\ x\ \neg p)$.

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

Proposition 3.22. $\models (\mathsf{GOAL}\, x\, p) \wedge (\mathsf{GOAL}\, x\, (p \supset q)) \supset (\mathsf{GOAL}\, x\, q)$.

Again, we have a necessitation property:

Proposition 3.23. *If* $\models \alpha$, *then* $\models (\mathsf{GOAL}\, x\, \alpha)$.

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

3.6.1. *Achievement goals*

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

Definition 3.24. $(\mathsf{A\text{-}GOAL}\, x\, p) \stackrel{\text{def}}{=} (\mathsf{GOAL}\, x\, (\mathsf{LATER}\, p)) \wedge (\mathsf{BEL}\, x\, \neg p)$.

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 goal, where what is desired for the future is something that is believed to be currently false.

3.6.2. *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 3.25. $\models \Diamond \neg (\mathsf{GOAL}\, x\, (\mathsf{LATER}\, p))$.

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);[14] but consider a plausible logical representation of that sentence in our formal language:

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

This sentence may be true, but it does not express an achievement goal since at some points the existential 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

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

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 3.25 when combined with an agent's commitments. First, we need to examine what, in general, are the consequences of having goals.

3.6.3. *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 3.26. $\models (\mathsf{BEL}\ x\ p) \supset (\mathsf{GOAL}\ x\ p)$.

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 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

[14] However, we have assumed immortal agents.

considers the future. Let p be of the form $\Diamond q$. From Proposition 3.26, 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 3.26 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:

$$(\mathsf{BEL}\ x \Diamond p \wedge \Box[\neg(\mathsf{BEL}\ x \Diamond p) \supset (\mathsf{GOAL}\ x \Box\neg p)]) .$$

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 3.26, agent's beliefs and goals "line up" with respect to their own primitive actions that happen next.

Proposition 3.27.

$$\models \forall x,e\ (\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ e)) \supset (\mathsf{GOAL}\ x\ (\mathsf{HAPPENS}\ x\ e)) .$$

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 3.28. *Expected consequences*:

$$\models (\mathsf{GOAL}\ x\ p) \wedge (\mathsf{BEL}\ x\ (p \supset q)) \supset (\mathsf{GOAL}\ x\ q) .$$

By Proposition 3.26, if an agent believes $p \supset q$ is true, $p \supset q$ is true in all his chosen worlds. Hence by Proposition 3.22, 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 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 towards 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 4.1.

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

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 3.26), 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 to achieve 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 is closed only under logical equivalence. First, we examine conjunction and disjunction. Then, we turn to implication.

4.1.1. *Conjunction, disjunction, and negation*

Proposition 4.2. *The logic of* P-GOAL:

$$\not\models (\text{P-GOAL}\ x\ (p \vee q)) \supseteq (\text{P-GOAL}\ x\ p) \wedge (\text{P-GOAL}\ x\ q) ,$$
$$\not\models (\text{P-GOAL}\ x\ (p \vee q)) \supseteq (\text{P-GOAL}\ x\ p) \vee (\text{P-GOAL}\ x\ q) ,$$
$$\models (\text{P-GOAL}\ x\ \neg p) \supset \neg(\text{P-GOAL}\ x\ p) .$$

First, (P-GOAL x (p ∧ q)) does not imply (P-GOAL x p) ∧ (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.[15] 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. But, 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) (i.e., ¬p ∧ ¬q).

With respect to the last property, note that while 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 3.11, (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 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 depends 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 it currently being true.

Table 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))	N (Y)
5	⊨ p ⊃ q	N (Y)
6	⊨ p ≡ q	Y

[15] For example, I may be committed to your knowing q, but not achieving q itself.

Consider Case 3, where the agent believes the implication *always* holds. Although Proposition 3.26 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 $p \supset q$ 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. Case 3 is, we believe, the norm.

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 towards 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 3.25). Hence, if an agent takes on a P-GOAL, he must give it up subject to the constraints imposed by P-GOAL.

Proposition 4.3. $\models (\text{P-GOAL x q}) \supset \Diamond[(\text{BEL x q}) \vee (\text{BEL x } \Box\neg q)]$.

This proposition is a direct consequence of Assumption 3.25, the definition of P-GOAL, and Proposition 3.8. 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. A simple consequence of Proposition 4.3 is:

Proposition 4.4.

$$\models \forall e(\text{P-GOAL x (DONE x e)}) \supset \Diamond[(\text{DONE x e}) \vee (\text{BEL x } \Box\neg(\text{DONE x e}))].$$

By Proposition 4.3, the agent eventually believes that he has done the act or that he will never do it. By Assumption 3.20, if the agent believes he has just done the act, then he has. We now give a crucial theorem:

Theorem 4.5. *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*:

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

Proof. By Proposition 4.3, 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 4.3, 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 3.8, 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 p. 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 unachieveable, 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 4.3 is one where q is (HAPPENS x a). Such a goal is one in which the agent's goal is that eventually the next thing that happens is his doing action a. Eventually, the agent believes either the next

action is his, or the agent eventually comes to believe he will never get the chance. 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 4.3, the agent will eventually believe (HAPPENS x a). By Proposition 3.27, 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 [8] 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 3.27 requires only that agents have expected effects as goals, but not 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 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 3.27, 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 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 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 their 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 that 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 satisfiable 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. Beliefs about Actions

We will define an agent's intending to do an action as that agent's forming a commitment, a P-GOAL to have done that action believing one is about to do it.

Before defining intention formally, we will need two assumptions regarding the beliefs an agent has about what he is about to perform.[16]

5.1. Belief and action

Assumption 3.20 characterizes retrospective beliefs about the past performance of primitive actions. Intentions will involve beliefs about complex actions that are about to be done. As usual, such beliefs about complex actions will be built upon beliefs about the agent's own primitive actions that are about to be done.

First, consider an agent's beliefs about what will be true after a sequence of actions he is about to do:[17]

Assumption 5.1.

$$\models \forall e\,(\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ e;\alpha?)) \supset (\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ e;(\mathsf{BEL}\ x\ \alpha)?))\ .$$

In other words, if an agent believes he is about to do e resulting in a world where α is true, then he also believes that after e, he will realize that α is true. Of course, in actuality, things could go awry, and he could change his beliefs after doing the action. But, what this assumption says is that he *now* believes that he will form the belief after e.[18] In particular, it follows from this assumption that if an agent believes that he is about to do two actions, the first being primitive, he also believes that after the first one, he will believe he is about to do the second, that is, it follows that (BEL x (HAPPENS x e;a)) implies (BEL x (HAPPENS x e;(BEL x (HAPPENS x a))?)).

The final assumption we make is that agents are not undecided about which, if any, event they believe they are about to do next. First, we adopt the following abbreviation:

– *Singleton sequence*:

$$(\mathsf{SINGLE}\ e) \stackrel{\text{def}}{=} (e \neq \mathsf{NIL}) \wedge (\forall x\,(x \leqslant e) \supset (x = e) \vee (x = \mathsf{NIL}))\ .$$

Singleton sequences are those that have only themselves and the empty sequence as subsequences.

The assumption is:

[16] We thank Joe Nunes for correcting and considerably simplifying an earlier version of the assumptions to follow.

[17] Actually, what counts is that an agent is about to do something *that is next* in the course of events describing the world. This limitation occurs because we are not considering simultaneous actions. Future work should loosen this restriction.

[18] This assumption would not hold with the event variable replaced by an arbitrary action expression. In the right circumstances, it is possible for an agent to think he is about to perform an iterative action without believing he will know when the termination condition is satisfied. See the discussion of iterative actions below.

Assumption 5.2.

$$\models \forall e\,(\mathsf{AGT}\,x\,e) \wedge (\mathsf{SINGLE}\,e) \supset$$
$$(\mathsf{BEL}\,x\,(\mathsf{HAPPENS}\,e)) \vee (\mathsf{BEL}\,x\,\neg(\mathsf{HAPPENS}\,e)).$$

In other words, for each single event of which x is the agent, either he believes the next thing to happen is his causing that event, or he believes it is not the next thing to happen. As with Assumption 3.20, this assumption does not hold when e is replaced by an arbitrary action expression. For example, an agent may believe neither (HAPPENS x e;p?) nor its negation, if he has no way of knowing whether or not p will be true. Moreover, the assumption needs to be limited to singleton sequences. Otherwise, (BEL x (HAPPEN x e;(e_1|e_2))), for example, would imply that one of (BEL x (HAPPEN x e;e_1)) or (BEL x (HAPPEN x e;e_2)) had to be true. This would have the very undesirable effect of requiring an agent to know (even before beginning) which branch he will take, a decision that the agent should be able to postpone until after the execution of e. But what we would like and what *does* indeed follow from Assumption 5.2 is that agents must know the *first step* that will be taken:

Proposition 5.3.

$$\models (\mathsf{BEL}\,x\,\exists e \neq \mathsf{NIL}\,(\mathsf{HAPPENS}\,x\,e)) \supset$$
$$\exists e'\,(\mathsf{SINGLE}\,e') \wedge (\mathsf{BEL}\,x\,(\mathsf{HAPPENS}\,x\,e')).$$

The antecedent would be true if the agent believed he was about to do a complex action (e.g., 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, he must have settled on the first step. Moreover, by Assumption 5.1, the agent also believes initially that if there are to be other steps beyond the first one, then after that first step, he will know the second step to take, and so on throughout the execution of the complex action.

With these assumptions, and given the expansion of complex action expressions in terms of the primitives, we can now complete the description of the consequences of an agent's believing he is about to do a complex action. First, consider disjunctive actions:

Proposition 5.4. *Agents are not nondeterministic*:

$$\models \forall e_1 \neq \mathsf{NIL}, e_2 \neq \mathsf{NIL}$$
$$(\mathsf{BEL}\,x\,(\mathsf{HAPPENS}\,x\,e_1|e_2)) \supset$$
$$(\mathsf{BEL}\,x\,(\mathsf{HAPPENS}\,x\,e_1)) \vee (\mathsf{BEL}\,x\,(\mathsf{HAPPENS}\,x\,e_2)) \vee$$
$$\exists z\,(\mathsf{SINGLE}\,z) \wedge (z \leqslant e_1) \wedge (z \leqslant e_2) \wedge (\mathsf{BEL}\,x\,(\mathsf{HAPPENS}\,x\,z)).$$

That is, if an agent believes he is about to do a disjunction of (sequences of) primitive events, then he must believe he is about to do one, or believe he is about to do the other, or believe he is about to do something that is common to both of them. For example, if an agent believes the next action in the world is his lifting his arm or his moving his foot, then the agent has an opinion on which act he will do. This is a consequence of Proposition 5.3.[19]

As a possible counterexample, imagine two agents, A and B having a fight. A believes he is about to block B's punch by either lifting his right or lifting his left arm. However, in our model, A does not believe that his blocking action is the next action; the next action is B's swinging. Once B swings, whichever act A does next will follow from A's beliefs (albeit quickly, and perhaps unconsciously). If, in fact there is no other intervening action (as with the example of the donkey placed between two bales of hay at equal distances) then nothing can change, so no decision will be made, and no action will take place.[20]

From Assumption 5.1, Proposition 5.4, and the definition of conditional actions, we can now show that an agent who is about to do a conditional action must believe its condition to be true, or believe it to be false. More generally, if he believes he will do a conditional action in the future, he believes he will have an opinion at the right time on the truth of the condition. Formally, one can show that:

Proposition 5.5. *If-then-else*:

$$
\begin{aligned}
\models\ & \forall e_1\ \forall e_2\ \forall e_3 \\
& [(\mathsf{SINGLE}\ e_2) \wedge (\mathsf{SINGLE}\ e_3) \wedge (e_2 \neq e_3) \wedge \\
& (\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ e_1;[\mathsf{IF}\ p\ \mathsf{THEN}\ e_2\ \mathsf{ELSE}\ e_3]))] \supset \\
& (\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ e_1;[(\mathsf{BEL}\ x\ (p \wedge (\mathsf{HAPPENS}\ x\ e_2))) \vee \\
& \qquad\qquad (\mathsf{BEL}\ x\ (\neg p \wedge (\mathsf{HAPPENS}\ x\ e_3)))]?))
\end{aligned}
$$

Essentially, the proof is this: In believing that $(\mathsf{HAPPENS}\ x\ e_1;[\mathsf{IF}\ p\ \mathsf{THEN}\ e_2\ \mathsf{ELSE}\ e_3])$, x believes $(\mathsf{HAPPENS}\ x\ e_1;(p?;e_2|\neg p?;e_3))$. Assumption 5.1 justifies our considering x as coming to the belief that $(\mathsf{HAPPENS}\ x\ (p?;e_2|\neg p?;e_3))$. By Proposition 5.4, x will believe he is about to do e_2 or he will believe he is about to do e_3. But, he believes he will only do e_2 if p holds, and e_3 otherwise. So, he must also come to the belief that p holds and he is about to do e_2, or $\neg p$ holds, and he is about to do e_3.

Now, the agent will have an opinion about which part of the conditional he will do *provided* that the *then* part and the *else* part do not share a common

[19] Notice also that although we have given semantics to nondeterministic actions, *agents* are themselves deterministic.

[20] This is unrealistic, of course. Ultimately, the passage of time is sufficient to change beliefs. Perhaps one way to accommodate this in future models is to treat the passage of time as a natural event.

first step (as in the above, where they are distinct events). A case where this would not hold is that of a physical "test," such as testing that a liquid is acidic or basic.[21] An agent who believes he is about to do:

$$\text{ACID?;Dip;(RED Paper)?}|\neg\text{ACID?;Dip;(BLUE Paper)?}$$

should not have to know in advance whether the liquid is acidic or basic. The third disjunct in Proposition 5.4 allows for this possibility.

Finally, turning to iterative actions, we have the following:

Proposition 5.6. *While-loops*:

$$\begin{aligned}
\models\ & \forall e_1\ \forall e_2 \\
& [(\text{SINGLE } e_1) \wedge (\text{SINGLE } e_2) \wedge (e_2 \neq e_1) \wedge \\
& \ (\text{BEL x (HAPPENS x [WHILE p DO } e_1];e_2))] \supset \\
& (\text{BEL x } \neg p) \vee \\
& (\text{BEL x } (p \wedge (\text{HAPPENS x } e_1;(\text{BEL x (HAPPENS x} \\
& \qquad\qquad [\text{WHILE p DO } e_1];e_2))?))) .
\end{aligned}$$

That is, if an agent believes he is about to do a while-loop, then he either believes that the condition is false (and does nothing) or believes it is true and that he is about to take one step of the loop, after which, he will be in the same state. As with the if-then-else, this holds *provided* the events in the while-loop are disjoint from any subsequent action (as in the above, where the event following the loop is distinct from the ones in the loop). A case where this would not be true is the following: Suppose an agent decides to repeat some action e a certain number of times believing that at some point in the sequence, perhaps at the very start, p will be false and remain so until the end. Even if the agent does not know exactly when p will be false, he nonetheless believes that he will do e at least until p is false. Thus, the agent believes he is about to do [WHILE p DO e], even though he does not know initially whether or not p is true. Moreover, in contrast to Assumption 5.1, he also believes that at the end of the while-loop, p will be false but that he may not realize it at the point. On the other hand, when the agent believes he has done a while-loop, he believes the condition is false.

The import of these analyses of belief and action is to show that agents can reason about complex actions (our analogue of plans) without having complete knowledge of how the world will unfold. Rather, one can acquire the needed information during the action's execution [36].

[21] See Moore [36] for another analysis of such tests.

At this point, we have characterized the dependencies among an agent's beliefs, the actions he has taken, and the actions he is about to take (that are next). These dependencies will be vital to an understanding of intention.

6. Intention as a Kind of Persistent Goal

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

6.1. INTEND_1

Typically, one intends to do actions. Accordingly, we define INTEND_1 to take an action expression as its argument.

Definition 6.1.

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

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 (i.e., 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.

Secondly, it is a commitment to success—to having done the action. As a contrast, consider the following inadequate definition of INTEND_1:

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

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.[22] 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.

6.1.1. *Intending actions*

Let us apply INTEND_1 to each kind of action expression. Recall that, Proposi-

[22] Notice that e could be the last step of a.

tion 4.3, intending to do an action results in the agent's eventually forming the belief that the action has been done (when the agent believed it was about to happen), or eventually believing it will never happen. Our interest here is in the former. The previous section discussed the consequences of believing one was about to do a complex action next.

First, consider intentions to "test" p. $(\mathsf{INTEND}_1\ x\ p?)$ expands into

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

By Proposition 3.1, this is equivalent to $(\mathsf{P\text{-}GOAL}\ x\ (\mathsf{DONE}\ x\ (\mathsf{KNOW}\ x\ p)?))$, which reduces to $(\mathsf{P\text{-}GOAL}\ x\ (\mathsf{KNOW}\ x\ p))$. 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 knocking down a tree:

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

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 will indeed knock down the tree. Notice that e is quantified outside of the 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

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

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 (i.e., 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 (i.e., here, he will believe that the tree is down).

Consider intending a conditional action. (INTEND$_1$ x [IF p THEN a ELSE b]) expands into

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

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. As we discussed in Proposition 5.5 the agent cannot believe he is about to do a conditional (more generally, a disjunction) 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.

Finally, consider intending sequences of actions. One can easily show that an agent who intends a;b intends to do a. However, at the start, that agent does *not* intend to do b. Rather, the agent intends to do (DONE agent a)?;b. That is, the agent intends to do b in the context of having just done a, but does not intend to do b by itself. However, once the agent believes he has just done a while executing a;b, the agent *then* intends b. All along, of course, the agent has the intention to do a;b, so that should the agent fail, he would be committed to trying again. Thus, intentions to do complex actions result in intentions at the right times to do the component actions.

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 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 INTEND$_1$. To allow for this case, we define another kind of intention, INTEND$_2$.

6.2. INTEND$_2$

One might intend to become rich, become happy, or (perhaps controversially) to kill one's uncle,[23] without having any idea how to achieve that state of affairs, not even having an enormous disjunction of possible options. In these cases, we shall 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.

[23] We are not trying to be morbid here; just setting up a classic example.

Definition 6.2.

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

We shall explain this definition in a number of steps. First, notice that to 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.[24] From Proposition 5.3 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 define what it means for an agent to have a plan.[25] The best we can do, and that is not too far off, is to say that agent must believe he 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.[26]

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, while 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 [11] and discussed in Searle's book [47]. 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 pedestrian, 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 INTEND_2, but without the condition stating that the agent should not want e not to bring about p. Call this

[24] 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 INTEND_1.

[25] See [41] for a discussion of these issues.

[26] 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.

$\text{INTEND}_{2'}$. So, assume the following is true: ($\text{INTEND}_{2'}$ x (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 $\text{INTEND}_{2'}$ 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, the agent not only believes he is about to do some sequence of events to bring about 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 3.26, 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.

7. Meeting the Desiderata for Intention

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 shall use INTEND_1 or INTEND_2 as best fits the example. Similar results hold for analogous problems posed with the other form of intention.

7.1. Bratman's functional roles played by intention

We reiterate Bratman's [7, 9] analysis of the roles that intentions typically play in the mental life of agents:

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. But, 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 shown in Section 6.1.1 that eventually, the agent will know which actions should be taken next. In the case of nonspecific intentions, such as $(\mathsf{INTEND}_2\ x\ p)$, we can derive via Proposition 4.3 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 non-specific intention, but does not encode the solution—that the agent will form a plan specifying just what that sequence of events would be.

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 7.1. *Screen of admissibility*:

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

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.[27]

Notice that our agents cannot knowingly (and hence, by Proposition 3.27, deliberately) act against their own best interests. That is, they cannot intentionally act in order to make their persistent goals unachieveable. 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.

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, and then does something, e, thinking it would bring about

[27] Notice that the theorem does not require quantification over primitive acts, but allows a and b to be arbitrary action expressions.

the doing of a, but he 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 7.2.

$$\models (\mathsf{DONE}\ x\ [(\mathsf{INTEND}_1\ x\ a) \wedge (\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ a))]?;e) \wedge (\mathsf{BEL}\ x\ \neg(\mathsf{DONE}\ x\ a)) \wedge \neg(\mathsf{BEL}\ x\ \Box\neg(\mathsf{DONE}\ x\ a)) \supset (\mathsf{INTEND}_2\ x\ a)\ .$$

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 unachieveable. Hence, the agent keeps the intention.

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

The agent believes it is possible to do action a

We do not have a modal operator for possibility. But we can state, via Proposition 3.26, 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.

Sometimes, the agent believes he will in fact do a

This is a consequence of Theorem 4.5, which states the conditions (call them *C*) under which $\Diamond(\mathsf{DONE}\ x\ a)$ holds, given the intention to do a. So, if the agent believes he has the intention, and believes *C* holds, $\Diamond(\mathsf{DONE}\ x\ a)$ follows from his beliefs as well.

The agent does not believe he will never do a

This principle is embodied directly in Proposition 3.26, 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).

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 3.28). 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.

At this point, the formalism captures each of Bratman's principles. Let us now see how it avoids the "Little Nell" problem.

7.2. Solving the "Little Nell" problem: When not to give up goals

Recall that in the "Little Nell" problem, Dudley never saves Nell because he believes he will be successful. Persistent goals avoid this problem; if Dudley adopts a persistent goal, he will drop it when he believes he has saved her, not when he believes he will save her. Thus, we have advocated the following initially plausible principles:

(1) Give up the intention that p when you believe p holds.
(2) Under at least some circumstances, an agent's intending that p entails the agent's believing that p will eventually be true.

These principles sound reasonable. One would think that a robot who forms the intention to bring someone a bottle of beer should drop that intention as soon as he brings the person the beer. Not dropping it constrains the adoption of other intentions, and may lead to person's receiving a year's supply of beer. The second principle says that at least in some (perhaps the normal) circumstances, one believes one's intentions will be fulfilled. Both of these principles can be found in our analysis.

One might think these principles entail a problem when one adopts a temporal or dynamic logic (modal or not, branching or not) that expresses "p will be true" as (FUTURE p) or $\Diamond$p. Apply principle (2) to a proposition p of the form $\Diamond$q. For example, let p represent "Nell is out of danger" by $\Diamond$(Saved Nell).[28] Hence, if the agent has the intention to bring about $\Diamond$(Saved Nell), under the right circumstances ("all other things being equal"), the agent believes $\Diamond\Diamond$(Saved Nell). But, in most temporal logics, $\Diamond\Diamond$(Saved Nell) entails $\Diamond$(Saved Nell). So, it is likely that the agent believes that $\Diamond$(Saved Nell) holds as well. Now, apply principle (1). Here, the agent had the intention to achieve $\Diamond$(Saved Nell) and the agent believes it is already true! So, the agent drops the intention, and Nell gets mashed.

[28] Notice that this goal is a bit out of the ordinary. The only way we can make sense of it is to allow Dudley to want to scare Nell a bit by, for example, letting her hear the approaching train, so that she will be much more grateful to her savior.

Our theory of intention based on P-GOAL avoids this problem because an agent's having a P-GOAL requires that the goal be true later and that the agent not believe it is currently true. The cases of interest are those in which the agent purportedly adopts the intention and believes he will succeed. Thus, $(\mathsf{BEL}\, x \Diamond\Diamond q)$, and so $(\mathsf{BEL}\, x \Diamond q)$. However, while it is certainly possible to intend to achieve q, an agent *never* forms the intention to achieve anything like $\Diamond q$ since, as already noted, $(\mathsf{P\text{-}GOAL}\, x \Diamond q)$ is always false.

One might argue that this analysis prevents agents from dropping their intentions when they think another agent will achieve the end goal. For example, one might want it to be possible for Dudley to drop the intention to save Nell (himself) because he thinks someone else, e.g., McDermott's Dick Daring, is going to save her. There are two cases to consider. The first case involves goals that can only be achieved once. The second case concerns goals that can be reachieved. We treat the second case in the next section. Regarding the first, one can easily show the following:

Theorem 7.3. *Dropping futile intentions*:

$$\begin{aligned}\models\ & \forall x\,(y \neq x) \wedge (\mathsf{BEL}\, x \Diamond \exists e\,(\mathsf{DONE}\, y\, \neg p?;e;p?)) \supset \\ & \neg(\mathsf{INTEND}_2\, x\, p) \vee \\ & \neg(\mathsf{BEL}\, x\,[\exists e\,(\mathsf{DONE}\, y\, \neg p?;e;p?) \supset \Box\neg\exists e\,(\mathsf{DONE}\, x\, \neg p?;e;p?)])\,.\end{aligned}$$

That is, if an agent believes anyone else is truly going to achieve p, then either the agent does not intend to achieve p himself, or he does not believe p can be achieved only once. Contrapositively, if an agent intends to achieve p, and always believes p can only be achieved once, the agent cannot simultaneously believe someone else is definitely going to achieve p. Intuitively, the reason is simple. If the agent believes someone else is definitely going to achieve p, then, because the agent believes that after doing so no one else could do so, the agent cannot have the persistent goal of achieving p himself; he cannot consistently believe he will achieve p first, nor can he achieve p later. A more rigorous proof is left to the determined reader.

Finally, our approach even allows Dudley to race Dick to save Nell. This is possible because in a true race, Dudley would not believe Dick will definitely win, and vice versa for Dick. Hence, Dudley would not be required to drop his intention. If Dudley did think Dick would definitely win, he might still run the race, but in pursuit of a different intention, for example, to finish the race, or convince Nell that he was trying, etc.

We have the intuition that Dudley should not drop his plan believing it will be successful because the only justification he has for that belief is his intending to do the planned actions. On the other hand, we should allow Dudley to drop his plan to save Nell when there is an alternate justification. McDermott [35]

advocates a data-dependency approach for recording such justifications. We cannot do likewise, but so far do not need to.

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.

8. An End to 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 our expanding the conditions under which one can drop a persistent goal.

8.1. Relativized persistent goal

Definition 8.1. *Persistent, relativized goal*:

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

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 $\neg$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 (e.g., [9, 15, 40]). Although we have not given a formal analysis of plans here (see [40] for such an analysis), the commitments one undertakes with respect to an action in a plan depend on the other planned actions, as well as the pre- and post-conditions 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 sufficient (but not necessary) prerequisite for his dropping a

subgoal.[29] 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 8.2.

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

Definition 8.3.

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

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:

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

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 decides to achieve p himself, relative to that persistent goal. So, the following is true of the agent:

$$(\text{P-GOAL}\ x\ p) \wedge (\text{INTEND}_2\ x\ p\ (\text{P-GOAL}\ x\ p))\ .$$

If someone else achieves p (and the agent comes to believe is true), the agent

[29] Also, notice that (P-GOAL x p) is now subsumed by (P-R-GOAL x p ¬p).

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 a 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 [14]).

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. This goes partway towards realizing a notion of "joint agency" espoused by Searle [48].[30]

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. It should be emphasized that a description of intention and its role in maintaining rational balance is not the same as a theory of intention formation. We are now in a position to study the latter.

8.2. Social commitments

Recently, a number of authors have emphasized the role that social commitments play in fostering and regulating social interaction. Fikes [17] points out

[30] 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 having adopted it relative to the second having adopted it, etc. For an analysis of circular propositions that might make such concepts expressible, see [5].

that office work is an informal domain whose functioning depends critically on the forming, negotiating, satisfying, monitoring, and discharging of workers' commitments to one another. For example, in committing to perform a service for someone, goal achievement (performing the service) is secondary to being released from the commitment. One way to be released is to convince the consumer of the service that the goal is achieved, is not achieveable, or is unnecessary. These are precisely the conditions for dropping a P-R-GOAL. Winograd and Flores [49] claim that social commitments constitute the foundation of communication and social interaction, and, by extension, office work. We agree with both of these claims. However, neither of these works provides an analysis of social commitment. This paper offers an analysis of what is, to us at least, a logically prior concept—an agent's having an *internal* commitment. It makes little sense to speak of social commitment if agents could not be internally committed.[31] We provide a framework for specifying conditions regulating when an agent can drop an internal commitment, conditions that map reasonably well onto Fikes' and Winograd and Flores' observations.

9. Implications for System Architecture

The theory presented here can be regarded as a specification for the design of an artificial agent, and should not be viewed as a logic with which an agent should reason. How, then, does this specification relate to system architecture?

First, the theory states that agents should have a distinguished mental state, call it GOAL, that is related in the ways we discussed to the mental state encoding the system's knowledge or beliefs (which we have termed BEL). As Rosenschein and Kaelbling [46] have shown, systems can be designed to satisfy knowledge (and belief) axioms without themselves manipulating knowledge (or belief) formulas.[32] Moreover, the mental states encoding the agent's beliefs and goals need to bear the appropriate relations to the agent's own primitive actions. We have described these relationships in some detail, but we have not been able to specify the *causally self-referential* connection between these mental states and the production of action. However, that should not be too surprising given the long-standing philosophical issues involved (e.g., see [26, 47]). The present theory merely constrains that causal connection. Apparently, it is easier to build systems that embody that causal connection than it is to describe it formally.

The second design principle to be gained from the present work is that agents should *be* committed. That is, their being in the state GOAL with respect

[31] We do not view young children as being socially committed until they have reached an "appropriate" state of maturity. One suggestion of our work is that that "appropriate" stage includes what we are terming "rational balance."

[32] The extension of their "situated automata" method of encoding knowledge to deal with incomplete knowledge of the mental states of other agents, is still an open problem.

to some propositional content should persist at least until the conditions specified herein for P-R-GOAL obtain. Of particular interest are the "relevance" conditions under which one would drop a relativized persistent goal. One could design a system that maintains rational balance by keeping its goals in a dependency network along with its beliefs, as has been done for the robot Flakey at SRI under the Intelligent Communicating Agents project. The robot would be in a goal state at least until the supporting mental states change. The relationship between specification and architecture is thus made apparent by encoding that dependency network in the last argument position of P-R-GOAL. Such agents will be committed in virtue of their architecture. But, although individual agents need not be built to reason explicitly about their *own* intentions and commitments, they will need to reason about the intentions and commitments of other agents in order to engage in communication. Thus, our developing a method by which agents reason about the intentions and commitments of others is still an important goal.

10. Conclusion

This paper establishes basic principles governing the rational balance among an agent's beliefs, actions, and intentions. Such principles provide 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 three characteristic functional roles played by intentions [8, 9], 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 seem 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 [14], we present a theory of speech acts that builds on the foundations laid here.

Much work remains. The action theory only allowed for possible worlds

consisting of single courses of events. Moreover, there were no truly alternative worlds, as would be necessary for a branching-time logic. Further developments should include basing the analysis on partial worlds/situations [6], and on temporal logics that allow for simultaneous actions [2, 20, 32]. Undoubtedly, the theory would be strengthened by the use of default and nonmonotonic reasoning.

Lastly, we can now allay the reader's fears about the mental state of the rationally unbalanced robot, Willie. If manufactured according to our principles, it is guaranteed that the problems described will not arise again; Willie will act on its intentions, not in spite of them, and will give them up when you say so. Of course, Willie is not yet very smart (as we did not say *how* agents should form plans), but he is determined.

ACKNOWLEDGEMENT

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

REFERENCES

1. J.F. Allen, A plan-based approach to speech act recognition, Tech. Rept. 131, Department of Computer Science, University of Toronto, Toronto, Ont. (1979).
2. J.F. Allen, Towards a general theory of action and time, *Artificial Intelligence* **23** (1984) 123–154.
3. J.F. Allen and C.R. Perrault, Analyzing intention in utterances, *Artificial Intelligence* **15** (1980) 143–178.
4. D. Appelt, *Planning English Sentences* (Cambridge University Press, Cambridge, 1985).
5. J. Barwise and J. Etchemendy, *The Liar: An Essay in Truth and Circularity* (Oxford University Press, New York, 1987).
6. J. Barwise and J. Perry, *Situations and Attitudes* (MIT Press, Cambridge, MA, 1983).
7. M. Bratman, Casteñada's theory of thought and action, in: J. Toberlin, ed., *Agent, Language, and the Structure of the World: Essays Presented to Hector-Neri Casteñada with his Replies* (Hackett, Indianapolis, IN, 1983) 149–169.
8. M. Bratman, Two faces of intention, *Philos. Rev.* **93** (1984) 375–405.
9. M. Bratman, *Intentions, Plans, and Practical Reason* (Harvard University Press, Cambridge, MA, 1987).
10. H.N. Casteñada, *Thinking and Doing* (Reidel, Dordrecht, Netherlands, 1975).
11. R.M. Chisholm, Freedom and action, in: K. Lehrer, ed., *Freedom and Determinism* (Random House, New York, 1966).
12. P.R. Cohen and H.J. Levesque, Speech acts and the recognition of shared plans, in: *Proceedings Third Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, Victoria, BC (1980) 263–271.
13. P.R. Cohen and H.J. Levesque, Speech acts and rationality, in: *Proceedings Twenty-third Annual Meeting Association for Computational Linguistics*, Chicago, IL (1985) 49–59.
14. P.R. Cohen and H.J. Levesque, Rational interaction as the basis for communication, in: P.R.

Cohen, J. Morgan and M.E. Pollack, eds., *Intentions in Communication* (MIT Press, Cambridge, MA, 1990).
15. P.R. Cohen and C.R. Perrault, Elements of a plan-based theory of speech acts, *Cognitive Sci.* **3** (1979) 177–212; reprinted in: B. Webber and N. Nilsson, eds., *Readings in Artificial Intelligence* (Morgan Kaufmann, Los Altos, CA, 1981) 478–495.
16. R. Fagin and J.Y. Halpern, Belief, awareness, and limited reasoning: Preliminary report, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985) 491–501.
17. R. Fikes, A commitment-based framework for describing informal cooperative work, *Cognitive Sci.* **6** (1982) 331–347.
18. R. Fikes and N.J. Nilsson, STRIPS: A new approach to the application of theorem proving to problem solving, *Artificial Intelligence* **2** (1971) 189–208.
19. M.P. Georgeff, Communication and interaction in multi-agent planning, in: *Proceedings AAAI-83*, Washington, DC (1983) 125–129.
20. M.P. Georgeff, Actions, processes, and causality, in: *Proceedings Workshop on Planning and Reasoning about Action*, Timberline, OR (1986).
21. M.P. Georgeff and A.L. Lansky, A BDI semantics for the procedural reasoning system, Tech. Note, Artificial Intelligence Center, SRI International, Menlo Park, CA (1986).
22. H.P. Grice, Meaning, *Philos. Rev.* **66** (1957) 377–388.
23. A. Haas, Possible events, actual events, and robots, *Comput. Intell.* **1** (2) (1985) 59–70.
24. J.Y. Halpern and Y.O. Moses, A guide to the modal logics of knowledge and belief, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985).
25. D. Harel, *First-Order Dynamic Logic* (Springer, New York, 1979).
26. G. Harman, *Change in View* (Bradford Books, MIT Press, Cambridge, MA, 1986).
27. K. Konolige, A first-order formalization of knowledge and action for a multiagent planning system, Tech. Note 232, Artificial Intelligence Center, SRI International, Menlo Park, CA (1980); also in: J.E. Hayes, D. Michie and Y.-H. Pao, eds., *Machine Intelligence* **10** (Ellis Horwood, Chichester, 1982).
28. K. Konolige, Experimental robot psychology, Tech. Note 363, Artificial Intelligence Center, SRI International, Menlo Park, CA (1985).
29. K. Konolige and N.J. Nilsson, Multiple-agent planning systems, in: *Proceedings AAAI-80*, Stanford, CA (1980).
30. L. Lamport, "Sometimes" is sometimes better than "not never", in: *Proceedings Seventh Annual ACM Symposium on Principles of Programming Languages* (1980) 174–185.
31. A.L. Lansky, Behavioral specification and planning for multiagent domains, Tech. Note 360, Artificial Intelligence Center, SRI International, Menlo Park, CA (1985).
32. A.L. Lansky, A representation of parallel activity based on events, structure, and causality, in: *Proceedings Workshop on Planning and Reasoning about Action*, Timberline, OR (1986).
33. H.J. Levesque, A logic of implicit and explicit belief, in: *Proceedings AAAI-84*, Austin, TX (1984).
34. J. McCarthy and P.J. Hayes, Some philosophical problems from the standpoint of artificial intelligence, in: B. Meltzer and D. Michie, eds., *Machine Intelligence* **4** (American Elsevier, New York, 1969).
35. D. McDermott, A temporal logic for reasoning about processes and plans, *Cognitive Sci.* **6** (1982) 101–155.
36. R.C. Moore, Reasoning about knowledge and action, Tech. Note 191, Artificial Intelligence Center, SRI International, Menlo Park, CA (1980).
37. L. Morgenstern, A first order theory of planning, knowledge, and action, in: J.Y. Halpern, ed., *Theoretical Aspects of Reasoning about Knowledge* (Morgan Kaufmann, Los Altos, CA, 1986).
38. C.R. Perrault, An application of default logic to speech act theory, in: P.R. Cohen, J. Morgan and M.E. Pollack, eds, *Intentions in Communication* (MIT Press, Cambridge, MA, 1990).

39. C.R. Perrault and J.F. Allen, A plan-based analysis of indirect speech acts, *Am. J. Comput. Linguistics* **6** (3) (1980) 167–182.
40. M.E. Pollack, Inferring domain plans in question answering, Ph.D. Thesis, Department of Computer Science, University of Pennsylvania, Philadelphia, PA (1986).
41. M.E. Pollack, Plans as complex mental attitudes, in: P.R. Cohen, J. Morgan and M.E. Pollack, eds., *Intentions in Communication* (MIT Press, Cambridge, MA, 1990).
42. V.R. Pratt, Six lectures on dynamic logic, Tech. Rept. MIT/LCS/TM-117, Laboratory for Computer Science, MIT, Cambridge, MA (1978).
43. J.S. Rosenschein, Rational interaction: Cooperation among intelligent agents, Ph.D. Thesis, Department of Computer Science, Stanford University, Stanford, CA (1986).
44. J.S. Rosenschein and M.R. Genesereth, Communication and cooperation, Tech. Rept. 84-5, Heuristic Programming Project, Department of Computer Science, Stanford University, Stanford, CA (1984).
45. S.J. Rosenschein, Plan synthesis: A logical perspective, in: *Proceedings IJCAI-81*, Vancouver, BC (1981) 331–337.
46. S.J. Rosenschein and L.P. Kaelbling, The synthesis of machines with provably epistemic properties, in: J.F. Halpern, ed., *Theoretical Aspects of Reasoning about Knowledge* (Morgan Kaufmann, Los Altos, CA, 1986) 83–98.
47. J.R. Searle, *Intentionality: An Essay in the Philosophy of Mind* (Cambridge University Press, New York, 1983).
48. J.R. Searle, Collective intentionality, in: P.R. Cohen, J. Morgan and M.E. Pollack, eds., *Intentions in Communication* (MIT Press, Cambridge, MA, 1990).
49. T. Winograd and F. Flores, *Understanding Computers and Cognition: A New Foundation for Design* (Ablex, Norwood, NJ, 1986).

Received October 1987; revised version received October 1988