Intention = Choice + Commitment[1]

Philip R. Cohen
Artificial Intelligence Center
and
CSLI[2]
SRI International
Menlo Park, CA 94025

Hector J. Levesque[3]
Department of Computer Science
University of Toronto
Toronto, Ontario,
Canada M5S 1A4

Abstract

This paper provides a logical analysis of the concept of intention as composed of two more basic concepts, choice (or goal) and commitment. By making explicit the conditions under which an agent can drop her goals, i.e., by specifying how the agent is *committed* to her goals, the formalism provides analyses for Bratman's three characteristic functional roles played by intentions [Bratman, 1986], 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.

By now, it is obvious to all interested parties that autonomous agents need to infer the intentions of other agents—in order to help those agents, hinder them, communicate with them, and in general to predict their behavior. Although intent and plan recognition has become a major topic of research for computational linguistics and distributed artificial intelligence, little work has addressed what it is these intentions are. Earlier work equated intentions with plans [Allen and Perrault, 1980, Cohen and Perrault, 1979, Schmidt *et al.*, 1978, Sidner and Israel, 1981], and recent work [Pollack, 1986] has addressed the collection of mental states agents would have in having a plan. However, many properties of intention are left out, properties that an observer can make good use of. For example, knowing that an agent is intending to achieve something, and seeing it fail, an observer may conclude that the agent is likely to try again. This paper provides a formal foundation for making such predictions.

[1]This research was made possible in part by a gift from the Systems Development Foundation, in part by support from the Natural Sciences and Engineering Research Council of Canada, and in part by support from the Defense Advanced Research Projects Agency under Contract N00039-84-K-0078 with the Naval Electronic Systems Command. The views and conclusions contained in this document are those of the authors and should not be interpreted as representative of the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency, the United States Government, or the Canadian Government. An expanded version of this paper appears in *Reasoning about Actions and Plans: Proceedings of the 1986 Workshop at Timberline Lodge*, (sponsored by AAAI), Morgan Kaufman Publishers, Inc., 1987.

[2]Center for the Study of Language and Information.

[3]Fellow of the Canadian Institute for Advanced Research.

I. Intention as a Composite

We model intention as a composite concept specifying what the agent has *chosen* and how the agent is *committed* to that choice. First, consider agents as choosing from among their (possibly inconsistent) desires those they want most.[4] Call what that follows from these chosen desires, loosely, goals. Next, consider an agent to have a *persistent goal* if she has a goal that she believes currently to be false, and that remains chosen at least as long as certain conditions hold. Persistence involves an agent's *internal* commitment over time to her choices.[5] In the simplest case, a "fanatic" will drop her commitment only if she believes the goal has been achieved or is impossible to achieve. Finally, intention is modelled as a kind of persistent goal—a persistent goal to do an action, believing one is about to do it.

Both beliefs and goals are modelled here in terms of possible worlds. Thus, our formalism does not deal with the actual chosen desires of an agent directly, but only with what is true in all *chosen worlds*, that is, worlds that are compatible with those desires. As usual, this type of coarse-grained model will not distinguish between logically equivalent goals (or beliefs). Moreover, we assume that these chosen worlds are all compatible with the beliefs of an agent, which is to say that if she has chosen worlds in which p holds, and she believes that p implies q, then she has chosen worlds in which q holds.

Despite these severe closure conditions, a crucial property of intention that our model *does* capture is that an agent may or may not intend the expected consequences of her intentions. Consider the case of taking a drug to cure an illness, believing that as a side-effect, one will upset one's stomach. In choosing to take the drug, the agent has surely deliberately chosen stomach distress. But that was not her intention; she is not committed to upsetting her stomach. Should she take a new and improved version of the drug that does not upset her stomach, all the better.[6] A system that cannot distinguish between the two cases is likely to be more of a hindrance than a help.

In the next sections of the paper we briefly develop elements of a formal theory of rational action, leading up to

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

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

[6]If the agent were truly committed to gastric distress, for instance as her indicator that the drug was effective, then if her stomach were not upset after taking the drug, she would ask for a refund.

a discussion of persistent goals. Then, we discuss the logic of persistent goal and define a notion of intention. We also extend the concept of a persistent goal to a more general one—one in which the dependencies of the agent's commitments can depend on arbitrary propositions. Finally, we lead up to a theory of rational interaction and communication by showing how agents can have interlocking commitments.

II. Elements of a Formal Theory

Below, we give an abbreviated description of the theory of rational action upon which we erect a theory of intention. Further details of this logic can be found in [Cohen and Levesque, 1987].

A. Syntax

The language we will use has the usual connectives of a first-order language with equality, as well as operators for the propositional attitudes and for talking about sequences of events: (BEL x p) and (GOAL x p) say that x has p as a belief or goal respectively; (AGT x e) says that x is the only agent of the sequence of events e; $e_1 \leq e_2$ says that e_1 is an initial subsequence of e_2; and finally, (HAPPENS a) and (DONE a) say that a sequence of events describable by an action expression a will happen next or just happened respectively. An action expression here is built from variables ranging over sequences of events using the constructs of dynamic logic: a;b is action composition; a|b is nondeterministic choice; p? is a test action; and finally, a^* is repetition. The usual programming constructs like IF/THEN actions and WHILE loops can easily be formed from these. We will use e as a variable ranging over sequences of events, and a and b for action expressions.

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.

B. Semantics

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

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

Although each world has a fixed, predetermined future, agents usually do not know which world they are in. Instead, some of these worlds are compatible with the agent's beliefs and goals. This is specified by means of two accessibility relations B and G. For a given agent x, $B(\sigma, x, n, \sigma^*)$ holds if the world σ^* is compatible with what x believes in world σ at point n (and similarly for G and goals). Turning this around, we could say that a G-accessible world is any course of events that an agent would be satisfied with, and that goals are just those propositions that are true in all such worlds (and analogously for beliefs).

Finally, to complete the semantic picture, we need a domain of quantification D that includes all people and finite sequences of events, and a relation Φ, which at every world and index point assigns to each k-place predicate symbol a k-ary relation over D. These sets, functions, and relations together make up a semantic structure.

Assume that M is a semantic structure, σ one of its possible worlds, n an integer, and v a set of bindings of variables to objects in D. We now specify what it means for M, σ, v, n to *satisfy* a wff p, which we write as $M, \sigma, v, n \models$ p. Because of formulas involving actions, this definition depends on what it means for a sequence of events described by an action 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:[7]

1. $M, \sigma, v, n \models$ (BEL x p) iff for all σ^* such that $B(\sigma, v(\mathsf{x}), n, \sigma^*)$, $M, \sigma^*, v, n \models$ p.
2. $M, \sigma, v, n \models$ (GOAL x p) iff for all σ^* such that $G(\sigma, v(\mathsf{x}), n, \sigma^*)$, $M, \sigma^*, v, n \models$ p.
3. $M, \sigma, v, n \models$ (AGT x e) iff $v(\mathsf{e}) = e_1 e_2 \ldots e_m$ and for every i, $Agt(e_i) = v(\mathsf{x})$. Thus x is the *only* agent of e.
4. $M, \sigma, v, n \models (\mathsf{e}_1 \leq \mathsf{e}_2)$ iff $v(\mathsf{e}_1)$ starts $v(\mathsf{e}_2)$.
5. $M, \sigma, v, n \models$ (HAPPENS a) iff $\exists m, m \geq n$, such that $M, \sigma, v, n[\![\mathsf{a}]\!]m$. That is, a describes a sequence of events that happens "next" (after n).
6. $M, \sigma, v, n \models$ (DONE a) iff $\exists m, m \leq n$, such that $M, \sigma, v, m[\![\mathsf{a}]\!]n$. That is, a describes a sequence of events that *just* happened (before n).

Turning now to the occurrence of actions, we have:

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 $\sigma(n + i) = e_i, 1 \leq i \leq m$. 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 $\exists k, n \leq k \leq m$, 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.
3. $M, \sigma, v, n[\![\mathsf{a \mid b}]\!]m$ 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.
4. $M, \sigma, v, n[\![\mathsf{p?}]\!]n$ iff $M, \sigma, v, n \models$ p. The test action, p?, involves no events at all, but occurs if p holds, or "blocks" (fails), when p is false.
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 $1 \leq i \leq k - 1$, $M, \sigma, v, n_i[\![\mathsf{a}]\!]n_{i+1}$.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 p is *satisfiable* if there is at least one M, world σ, index n, and assignment v such that $M, \sigma, v, n \models$ p. A wff

[7] For conciseness, we omit that part of the definition that deals with the constructs of first-order logic with equality.

p is *valid*, iff for every M, world σ, event index n, and assignment of variables v, $M, \sigma, v, n \models$ p.

We will adopt the following abbreviations:

Actions: (DONE x a) $\stackrel{\text{def}}{=}$ (DONE a) $\wedge$ (AGT x a) and
(HAPPENS x a) $\stackrel{\text{def}}{=}$ (HAPPENS a) $\wedge$ (AGT x a).

Eventually: $\Diamond$p $\stackrel{\text{def}}{=}$ $\exists$e (HAPPENS e;p?).
$\Diamond$p is true if there is something that happens (including the null action) after which p holds, that is, if p is true at some point in the future.[8]

Later: (LATER p) $\stackrel{\text{def}}{=}$ $\sim$p $\wedge$ $\Diamond$p.

Always: $\Box$p $\stackrel{\text{def}}{=}$ $\sim\Diamond\sim$p.
$\Box$p means that p is true throughout the course of events from now on.

Before: (BEFORE p q) $\stackrel{\text{def}}{=}$
$\forall$c (HAPPENS c;q?) $\supset$ $\exists$a (a $\leq$ c) $\wedge$ (HAPPENS a;p?).
The wff p will become true no later than q.

Know: (KNOW x p) $\stackrel{\text{def}}{=}$ p $\wedge$ (BEL x p).

Competence: (COMPETENT x p) $\stackrel{\text{def}}{=}$ (BEL x p) $\supset$ p.
Agents that are competent with respect to some proposition have only correct beliefs about it.[9]

C. Properties and Assumptions

It is not too difficult to establish that action expressions as defined here have their dynamic logic interpretation. For example,

$\models$ (HAPPENS p?;(b|c)) $\equiv$ [p $\wedge$ ((HAPPENS b) $\vee$ (HAPPENS c))].

So a test action followed by a nondeterministic action happens iff the test is true and one of the two actions happens next. Moreover, HAPPENS and DONE interact, as in

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

So, for example, if an action happens next, then immediately afterwards, it is true that it just happened.

Note that there is a sharp distinction between action expressions and primitive events. Examples of the latter might include moving an arm, exerting force, and uttering a word or sentence. Action expressions are used to characterize sequences of primitive events that satisfy certain properties. For example, a movement of a finger may result in a circuit's being closed, which may result in a light's coming on. We will say that one primitive event happened, which can be characterized by various complex action expressions.

Turning now to the attitudes, they can be shown to satisfy the usual closure conditions:

$\models$ (BEL x p) $\wedge$ (BEL x (p $\supset$ q)) $\supset$ (BEL x q).
If $\models$ p then $\models$ $\Box$(BEL x $\Box$p).

(and similarly for GOAL). In addition we make the following assumptions:[10]

Agents Know: $\models$ (HAPPEN x e) $\supset$ (BEL x (HAPPEN e)).
A primitive event performed by an agent will occur only if its agent realizes it will. Accidental or unanticipated events are possible, but these are considered to happen *to* an agent. Note that this assumption does not apply to arbitrary action expressions here, since an agent may obviously achieve some state of affairs unknowingly.

Consistency: $\models$ (GOAL x p) $\supset$ $\sim$(GOAL x $\sim$p).
There is always at least one world compatible with the goals of an agent. Because of realism below, this also applies to belief.

Realism: $\models$ (BEL x p) $\supset$ (GOAL x p).
Every chosen world is compatible with an agent's beliefs. This is not to say that an agent cannot simultaneously believe that p is false and want p to be true at some later point; however, if an agent (that does not engage in wishful thinking) believes that p is false *now*, her chosen worlds all reflect this fact.

No infinite deferral: $\models$ $\Diamond\sim$(GOAL x (LATER p)).
Agents eventually drop all "achievement" goals—goals they believe are currently false but want to be true later. These either become "maintenance" goals—goals the agent believes are currently true and need only be kept true—or are dropped completely (for example, if the agent comes to believe they are unachievable).

Together, these assumptions imply that achievement goals must be consistent, compatible with all beliefs about the future, and of limited duration.

At this point, we are finished with the foundational level, having briefly described agents' beliefs and goals, events, and time. Further discussion can be found in [Cohen and Levesque, 1987].

III. Persistent Goals

To capture *one* grade of commitment (fanatical) that an agent might have toward her goals, we define a persistent goal, P-GOAL, to be one that the agent will not give up until she thinks it has been satisfied, or until she thinks it will never be true.[11] Specifically, we have

Definition 1 (P-GOAL x p) $\stackrel{\text{def}}{=}$
(GOAL x (LATER p)) $\wedge$ (BEL x $\sim$p) $\wedge$
(BEFORE [(BEL x p) $\vee$ (BEL x $\Box\sim$p)] $\sim$(GOAL x (LATER p))).

Notice the use of LATER, and hence $\Diamond$, above. P-GOALs are achievement goals; the agent's goal is that p be true in the future, and she believes it is not currently true. As soon as the agent believes it will never be true, we know the agent must drop her goal (by Realism), and hence her persistent goal. Moreover, as soon as an agent believes p is true, the belief conjunct of P-GOAL requires that she drop the persistent goal that p be true. 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 her *simple* goal when she thinks it is satisfied, since agents may have goals of maintenance.

[8] Note that $\Diamond$p and $\Diamond\sim$p are jointly satisfiable.

[9] It is reasonable to assume that agents are competent with respect to their own beliefs, goals, and their having done primitive events.

[10] In other words, we only deal with semantic structures where these propositions come out true.

[11] The latter case could arise easily if the proposition is one that specifically mentions a time (which for simplicity, we have not included in the formalism here). Once the agent believes that the time is past, she believes the proposition is impossible to achieve.

Thus, achieving one's persistent goals may convert them into maintenance goals.

A. The Logic of P-GOAL

The logic of P-GOAL is weaker than one might expect. We have the following:

1. $\not\models$(P-GOAL x p∧q) $\begin{matrix}\supset\\ \subset\end{matrix}$ (P-GOAL x p)∧(P-GOAL x q)
2. $\not\models$(P-GOAL x p∨q) $\begin{matrix}\supset\\ \subset\end{matrix}$ (P-GOAL x p)∨(P-GOAL x q).
3. $\models$(P-GOAL x ∼p) ⊃ ∼(P-GOAL x p)

First, (P-GOAL x p∧q) does not imply (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.[12] 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. Similar analyses can be given for the other properties of P-GOAL.

We now give a crucial theorem:

Theorem 1 *From persistence to eventualities—If someone has a persistent goal of bringing about* p, p *is always within her area of competence, and the agent will only believe that* p *will never occur after she drops her goal, then eventually* p *becomes true:*

$\models$ (P-GOAL y p) ∧ □ (COMPETENT y p) ∧
∼[BEFORE (BEL y □∼p) ∼(GOAL x (LATER p))] ⊃ ◇p.

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 she could forever create incorrect plans. If the goal is not persistent, we also cannot conclude ◇p since she could give up the goal without achieving it. If the goal actually is impossible for her to achieve, but she does not know this and commits to achieving it, then we know that eventually, perhaps after trying hard to achieve it, she will come to believe it is impossible and give up.

B. Relativized Persistent Goals

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

Definition 2 (P-R-GOAL x p q) $\stackrel{\text{def}}{=}$
(GOAL x (LATER p)) ∧ (BEL x ∼p) ∧
(BEFORE [(BEL x p) ∨ (BEL x □∼p) ∨ (BEL x ∼q)]
∼(GOAL x (LATER p))).

That is, a necessary condition to giving up a P-R-GOAL is that the agent believes it is satisfied, or believes it is impossible to achieve, or believes ∼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, an agent could adopt the persistent goal to buy an umbrella relative to her belief that it will rain. That agent could consider dropping her persistent goal should she come to believe that the forecast has changed.

One can prove a theorem analogous to Theorem 1: If someone has a persistent goal of bringing about p, relative to q, and, before dropping her goal, p remains within her area of competence, and the agent will not believe that p will never occur or believe that q is false, then eventually p becomes true.

At this point, we are ready to define intention. There are two forms of intention—intending actions and intending to achieve some state of affairs. For this brief paper, we only present the former; see [Cohen and Levesque, 1987] for the latter.

C. Intention Defined

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

Definition 3 (INTEND_1 x a) $\stackrel{\text{def}}{=}$
(P-GOAL x [DONE x (KNOW x (HAPPENS a))?;a]).

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 she is about to do the intended action, and then doing it.

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

(INTEND_1 x a) $\stackrel{\text{def?}}{=}$ (P-GOAL x (KNOW x (HAPPENS x a))).

This would say that an intention is a commitment to being *on the verge* of doing a (knowingly). 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.

Just as we refined our analysis of persistent goal to allow the commitment to be relative to the agent's believing arbitrary states-of-affairs, so too can we extend the above definition of intention:

Definition 4 (INTEND_1 x a q) $\stackrel{\text{def}}{=}$
(P-R-GOAL x [DONE x (KNOW x (HAPPENS a))?;a] q).

IV. Properties of Intentions

In this section we show how various properties of the commonsense concept of intention are captured by our analysis based on P-GOAL. First, we consider how our definitions characterize the functional roles that intentions are thought to play in the mental lives of agents [Bratman, 1984, Bratman, 1986]

1. *Intentions normally pose problems for the agent; the agent needs to determine a way to achieve them.* If the agent intends an action as described by an action expression, then she knows in general terms what to do. However, the action expression may have disjunctions or conditionals in it. Hence, she need not know at the time of forming the intention exactly what will be done. But unless she comes to believe the action

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

is unachievable, she must sooner or later correctly believe that she is about to do something that accomplishes the action (by Theorem 1). Now by the Knowing Agents assumption, a primitive event will occur only if its agent believes it will. So sooner or later, the agent will decide on a specific thing to do. Thus, agents are required to convert non-specific intentions into specific choices of primitive events to that end.

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

   Theorem 2
   $\models (\mathsf{INTEND}_1\ x\ b) \wedge \Box(\mathsf{BEL}\ x\ [(\mathsf{DONE}\ x\ a) \supset \Box\sim(\mathsf{DONE}\ x\ b)]) \supset \sim(\mathsf{INTEND}_1\ x\ a;b).$

   Thus our agents cannot intentionally act to make their persistent goals unachievable. For example, if they have adopted a time-limited intention, they cannot intend to do some other act knowing it would make achieving that time-limited intention forever false.

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

   Theorem 3
   $\models (\mathsf{BEL}\ x \sim(\mathsf{DONE}\ x\ a)) \wedge \sim(\mathsf{BEL}\ x \Box\sim(\mathsf{DONE}\ x\ a)) \wedge (\mathsf{DONE}\ x\ [(\mathsf{INTEND}_1\ x\ a) \wedge (\mathsf{BEL}\ x\ (\mathsf{HAPPENS}\ x\ a))]?;b) \supset (\mathsf{INTEND}_1\ x\ a).$

   Because an agent cannot give up an intention until it is believed to have been achieved or to be unachievable, the agent here keeps the intention.

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

4. *The agent does not believe she will never do* a. This principle is embodied directly in the assumptions of Consistency and Realism. If an agent forms the intention to do a, then in her chosen worlds, she eventually does a. But this is not realistic if she believes she will never do a.
5. *The agent believes that* a *can be done.* We do not have a modal operator for possibility, but we do have the previous property which may be close enough for current purposes.
6. *Sometimes, the agent believes she will in fact do* a. This is a consequence of Theorem 1, which states conditions under which a P-GOAL will eventually come to be true. So given that an agent believes both that she has the intention to do a and that these conditions hold, she will also believe $\Diamond(\mathsf{DONE}\ x\ a)$.
7. *Agents need not intend the expected side-effects of their intentions.* Recall that in an earlier problem, an agent intended to cure an illness believing that the necessary medicine would upset her stomach. The fact that the agent knowingly chooses to upset her stomach without intending to do so is accomodated in our scheme since $(\mathsf{INTEND}_1\ x\ a;p?) \wedge (\mathsf{BEL}\ x\ \Box(p \supset q))$ does not imply $(\mathsf{INTEND}_1\ x\ a;q?)$. The reason is that although there is a belief that p is inevitably accompanied by q, this *belief* could change over time (for example, if the agent finds out about new medicine). Under these circumstances, although p remains a persistent goal, q can now be realistically dropped. Thus, q was not a truly persistent goal after all, and so there was no intention.

   However, with $\Box(\mathsf{BEL}\ x\ \Box(p \supset q))$ as the initial condition, q can no longer be dropped, and so our formalism now says that q is intended. But this is as it should be. If the agent always believes, no matter what, that stomach upset is required by effective treatment, then in her commitment to such treatment, she will indeed be committed to upsetting her stomach, and track her attempts at that, just like any other intention.

We can also demonstrate that our notion of intention avoids McDermott's "Little Nell" problem [McDermott, 1982], in which an agent drops her intention precisely because she believes it will be successful. The problem can occur with any concept of intention (like ours) that satisfies the following two plausible principles:

1. An intention to achieve p can be given up when the agent believes that p holds.
2. Under some circumstances, an intention to achieve p is sufficient for the agent to believe that p will eventually be true.

The problem is when the intention p is of the form $\Diamond q$. By the second principle, in some cases, the agent will believe that eventually $\Diamond q$ will be true. But $\Diamond\Diamond q$ is equivalent to $\Diamond q$, and so, by the first principle, the belief allows the intention to be given up. But if the agent gives it up, $\Diamond q$ need not be achieved after all!

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. In particular, an agent *never* forms the intention to achieve anything like $\Diamond q$: because $(\mathsf{LATER}\ \Diamond q)$ is always false, so is $(\mathsf{P\text{-}GOAL}\ x\ \Diamond q)$.

Finally, our analysis supports the observation that intentions can (loosely speaking) be viewed as the contents of plans (e.g., [Bratman, 1986, Cohen and Perrault, 1979, Pollack, 1986]). Although we have not given a formal analysis of plans here (see [Pollack, 1986] 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 $(\mathsf{GOAL}\ x\ q)$, then necessary conditions for x's dropping her goal include her believing that she no longer has q as a goal. Thus, $(\mathsf{P\text{-}R\text{-}GOAL}\ x\ p\ (\mathsf{GOAL}\ x\ q))$ characterizes an agent's having a persistent *subgoal* p relative to the *supergoal* q. An agent's dropping a supergoal is now a necessary (but not sufficient) prerequisite for her dropping a subgoal. 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/intentions.

V. Conclusion

Autonomous agents need to be able to reason not only about the plans that other agents have, but also about their state of commitment to those plans. If one agent finds out that another has failed in attempting to achieve something, the first should be able to predict when the other will try again. The first agent should be able to reason about the other agent's intentions and commitments rather than be required to simulate the other agent's planning and replanning procedures.

This research has developed a formal theory of intention that shows the intimate relationship of intention to commitment. Whereas other logics have related belief and knowledge to action, we have explored the consequences of adding another modality for goals, and have examined the effects of keeping goals over time. The logic of intention derives from this logic of persistent goal, and is finer-grained than one might expect from our use of a possible-worlds foundation. It provides a descriptive foundation for reasoning about the intentions of other agents, without yet making a commitment to a reasoning strategy. Finally, it serves as the foundation for a theory of speech acts and communication [Cohen and Levesque].

VI. Acknowledgements

James Allen, Michael Bratman, Jim des Rivières, Joe Halpern, David Israel, Joe Nunes, Calvin Ostrum, Ray Perrault, Martha Pollack, and Moshe Vardi provided many valuable suggestions. Thanks to you all.

References

[Allen and Perrault, 1980] J. F. Allen and C. R. Perrault. Analyzing intention in dialogues. *Artificial Intelligence*, 15(3):143–178, 1980.

[Bratman, 1984] M. Bratman. Two faces of intention. *The Philosophical Review*, XCIII(3):375–405, 1984.

[Bratman, 1986] M. Bratman. Intentions, plans, and practical reason. 1986. Harvard University Press, in preparation.

[Cohen and Levesque] P. R. Cohen and H. J. Levesque. Rational interaction as the basis for communication. In preparation.

[Cohen and Levesque, 1987] P. R. Cohen and H. J. Levesque. *Persistence, Intention, and Commitment.* Technical Report 415, Artificial Intelligence Center, SRI International, Menlo Park, California, February 1987. Also appears in *Proceedings of the 1986 Timberline Workshop on Planning and Practical Reasoning*, Morgan Kaufman Publishers, Inc. Los Altos, California.

[Cohen and Perrault, 1979] P. R. Cohen and C. R. Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212, 1979. Reprinted in *Readings in Artificial Intelligence*, Morgan Kaufman Publishing Co., Los Altos, California, B. Webber and N. Nilsson (eds.), pp. 478-495., 1981.

[McDermott, 1982] D. McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6(2):101–155, April-June 1982.

[Pollack, 1986] M. E. Pollack. *Inferring Domain Plans in Question Answering.* PhD thesis, Department of Computer Science, University of Pennsylvania, 1986.

[Schmidt *et al.*, 1978] C. F. Schmidt, N. S. Sridharan, and J. L. Goodson. The plan recognition problem: an intersection of artificial intelligence and psychology. *Artificial Intelligence*, 10:45–83, 1978.

[Sidner and Israel, 1981] C. Sidner and D. Israel. Recognizing intended meaning and speaker's plans. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Vancouver, B. C., 1981.