Intentions in Time

Robert P. Goldman and R. Raymond Lang
Computer Science Department, 301 Stanley Thomas Hall, Tulane University
New Orleans, Louisiana 70118, (504) 865-5840
rpg@cs.tulane.edu and lang@cs.tulane.edu

Abstract

Representing and reasoning about goal-directed actions is necessary in order for autonomous agents to act in or understand the commonsense world. This paper provides a formal theory of intentional action based on Bratman's characterization of intention [Bratman, 1987, Bratman, 1990]. Our formalization profits from the the formalization of Bratman's theory developed by Cohen and Levesque [1990a, 1990b]. We review their formalization and illustrate its weaknesses. Using Allen's temporal logic [Allen, 1984] , we construct a formalization that satisfies Bratman's desiderata for an acceptable theory of intentional action. We introduce a characterization of success and failure of intentional action and show that our richer theory of time allows us to formalize more complex intentional actions, particularly those with deadlines. Finally, we argue that the use of a syntactic theory of belief allows us to accommodate a more descriptive theory of intentional action by fallible agents. Our work has relevance to multi-agent planning, speech-act processing and narrative understanding. We are using this theory to representing the content of narratives and to constructing and understanding description-based communication.

Introduction

This paper provides a formal theory of intentional action within a temporal logic. Autonomous systems acting in a commensense world need to be able to understand goal-directed actions. This ability is necessary for applications like multi-agent planning, speech-act processing and narrative understanding. In particular, our systems must be able to recognize the successes and failures of intentional actions and the ramifications of these outcomes. In this paper we present a formal theory of intentional action to meet these goals.

Our formalization is based on Bratman's characterization of intention [Bratman, 1987, Bratman, 1990]. Bratman argues that intention plays an important functional role and that rationality imposes constraints of consistency and feasibility on intentions.

Our formalization profits from the insights developed by Cohen and Levesque [1990a, 1990b] in their formalization of Bratman's theory. Cohen and Levesque (henceforth C&L) use dynamic logic [Harel, 1979] with a possible world semantics as a foundation for their system. While C&L do a good job of formalizing Bratman's theory, the dynamic logic framework has an impoverished notion of time and is severely limited in its expressiveness.

Allen [1984] has proposed a temporal logic intended for reasoning about events described in natural language. His logic is first-order with a semantics based on intervals over a single time line. Allen suggests that the logic be augmented by a quotation mechanism (syntactic theory) [Haas, 1986] to handle belief and intention contexts.

In this paper, we construct a formalization that, like C&L's, satisfies Bratman's desiderata for an acceptable theory of intentional action. We introduce a characterization of success and failure of intentional action. We show that our richer theory of time allows us to formalize more complex intentional actions, particularly those with deadlines. Finally, we argue that the use of a syntactic theory of belief will allow us to accommodate a more descriptive theory of intentional action by fallible agents. Thus we construct a more powerful theory than C&L's while avoiding the complications of modal logic.

Intention and Action

We take as our starting point Bratman's analysis of intention [Bratman, 1987, Bratman, 1990]. We distill from that analysis the following functional roles filled by intention:

1. Intentions serve as a sort of "top level plan" for the agent; the agent needs to determine how to carry them out.
2. Intentions limit an agent's further intentions to those consistent with ones already held.
3. An intention provides an agent with an indication of which features of the world he should "track" in order to determine the success of attempt(s) to achieve

the intention.

4. An agent is able to make predictions about the behavior of other agents based on what is known about their intentions.

In addition, Bratman goes on to say that intention should satisfy the following properties. If an agent intends an action a, then:

1. The agent believes a is a feasible and applicable means of achieving some goal p.
2. The agent does not believe it will bring about a state in which a is no longer feasible or in which the goal toward which a is directed is unachievable.
3. Under certain conditions, the agent believes it will do a.
4. Agents need not intend all the expected side effects of their intentions.

Following Bratman's analysis, C&L[1990a, 1990b] develop a set of constraints on rational agents such that a robot designed according to these constraints will conform to the following:

- It does not procrastinate forever (i.e, the robot will act on its intentions). Note that the link is from *intentions* to actions, not from *goals* to actions.
- The agent is persistent in the pursuit of its goals, subject to constraints of rationality.
- The agent drops goals when it determines that the goal need not be achieved, whether because the goal holds, is impossible, or the condition relative to which the goal was adopted no longer holds.

C&L base their account on the modal predicates `believes` and `goal`. Because C&L give a possible world semantics to the predicate, `goal` applies to all those propositions which are true in all worlds consistent with the explicit goals of the agent. If a proposition also is believed by the agent currently to be false, then the propostion is an `a-goal`, or "achievement goal." Further, if an agent holds an `a-goal` until the proposition is true or is believed by the agent to be impossible, then that goal is a `p-goal`, or persistent goal.

Intentions in C&L are p-goals of a certain form. Specifically, an intention is a p-goal that an action will have been done by the intending agent. The fulfillment of such a goal is precisely the doing of the action, so C&L are able to draw some very appealing conclusions from their definitions.

One drawback to C&L's reliance on the possible worlds approach to characterize goals is that every inevitable future event is a persistent goal. For example, let us consider an agent which knows that it will eventually die. So the agent is eventually dead in all worlds believed. Goal worlds are a subset of believed worlds. Ergo it is our agent's persistent goal to die.

Another problem with C&L's formalization of goals is that it provides no link between the possession of a goal and the performance of actions to bring that goal about. The link between *intentions* and actions in C&L is due to the definition of an intention as a persistent goal the content of which is some action. But C&L's reductionist formalization specifies only that a persistent goal will eventually be either (believed to be) satisfied or (believed to be) unachievable. For example, in their scheme, if one believes that one has successfully convinced another agent to help satisfy some goal proposition P, then one believes that this helper has a persistent goal (relative to one's persistent goal that P) to bring about P. From this and the reasonable, but defeasible, assumptions that the helper is competent to observe P and doesn't come to believe that it is impossible for P to be true, one may conclude that P will eventually become true. This seems like a Good Thing, but of course one would like to be justified in assuming — all else being equal — that one's helper will act to provide assistance in achieving the goal proposition. Unfortunately, a "helper" who simply waited while one achieved the goal oneself would satisfy C&L's specifications.

Allen's Temporal Logic

We base our formalization on Allen's logic of action and time [Allen, 1984].[1] Allen's logic is an interval-based first-order system whose ontology contains entities (objects and agents), properties and events, and quoted logical expressions. The temporal intervals of Allen's logic are intervals over a dense time line, and Allen provides relations to represent orderings between them. Events and properties denote sets of intervals: those over which events OCCUR and those over which properties HOLD, respectively. Typically, one describes event and property *types*, using predicate calculus functions. For example, `(falls fred)` denotes the set of intervals over which it OCCURs that Fred falls; `(on blocka blockb)` denotes the set of intervals when it HOLDs that block a is on block b.

Since Allen's logic is first-order, agents' beliefs are represented syntactically, as quoted strings. We feel that syntactic theories provide a more appealing model of belief for AI. First, the syntactic model of belief is also the strong AI model of intelligence: that reasoning can be represented in terms of symbolic manipulation. Furthermore, the syntactic theory offers us a way of avoiding the problem of consequential closure which plagues modal theories. We can write axioms which describe limited inferencing processes (see [Haas, 1986] for an example). Finally, whenever possible, we prefer to remain within the simpler first-order framework.

Notational conventions: In the following, we use a lisp-like, prefix notation for the predicate calculus. For the sake of brevity, we omit universal quantifiers; unbound variables should be assumed to be universally

[1] A very similar logic is provided by McDermott [1982]

quantified. We also use typed variables in an informal way: variable names indicate specific types of entities as described in Figure 1. In order to make the description of quoted expressions less cumbersome, we assume a backquote, or quasiquote operator which acts as its equivalent in Lisp. We also assume that all agents refer to themselves using the constant `self`, and refer to the present using the constant `now`.

a,b action terms of the form *(acause x e)*, where *x* is an agent and *e* is an event term as described by Allen in [1984]. Note that actions are a subclass of events.

e,f event terms

p,q strings denoting goal propositions.

s,t time intervals

x agents

$\mathbf{s_n}$ symbols

Figure 1: Types denoted by variable letters

Goals

A theory of intentional actions is necessarily a theory of *goal-directed* actions. Thus goals are a central component of a theory of intention. We argue that for an agent to have a goal is to have in mind a proposition about the world that it would like to make true. Accordingly, we argue that a statement describing a goal must minimally contain a HOLDS predication: goals must be directed towards the establishment of a fluent (changeable property). For the sake of simplicity, we do not allow events to be goals. Finally, we would like to be able to represent constraints on the time at which a state of affairs is to hold, in order to reason about deadlines, etc. This can be achieved by conjoining HOLDS predications with temporal constraints.

Accordingly, we provide a `goal` predicate of three arguments:

`(goal` *x p t*`)`

where *x* is a term denoting an agent, *p* is a quoted expression containing at least one HOLDS predication and *t* is a term denoting the interval over which *p* is held by *x* as a goal.

Note that our treatment of goals is very different from that of Allen (see [1984, p. 145]). Allen's goal predicate is to be used in expressions of the following form:

`(is-goal-of` *agent goal gtime t*`)`

which states that *agent* would like to make the property *goal* HOLD at time *gtime*. *t* is the interval over which the agent has this goal.

We wish to stress that for Allen a goal is a *property* rather than a string describing a state of affairs. There are two problems that arise when such a representation is used. First, there is a problem with Leibniz's law (referential transparency): since a property denotes a set of intervals, we must be able to substitute for one goal another property expression which is true over the same intervals. In particular, all properties which never hold are the same. Davis[1990, p. 414] gives a similar critique.

Second, the logic is unable to represent vague goals: this problem arises because we are unable to quantify in the expression describing a goal. Therefore, we are unable to formulate simple desires like "I want to own a red car someday." The best one can do in Allen's logic is

```
(exists x gt (goal me (AND (owns me x)
                           (red-car x))
                     gt now))
```

which says that there is some time and some red car which I want to own. The difficulty with this formalization is that two meanings are possible: (1) the agent wants some car at some time, but doesn't know which car or exactly when, or (2) the *agent* knows exactly which car he wants and when, but *we* don't.

Our goal representation is also more expressive than C&L's scheme. C&L's goals are restricted to achievement goals: an agent is described as wanting to bring about a state of affairs, which is not the case at present, at some indefinite time in the future. As a consequence of their use of dynamic logic, it is impossible for them to express either deadlines for goals or establishment times (times before which it is too soon). Furthermore, they regard it as illegitimate for an agent to have as its goal some state in the future which holds now. This would make it impossible for the owner of a pennant-winning ball club to intend to win the pennant next year.

We axiomatize the following constraints on goals:

- if *p* is an agent's goal at time *t*, then *not p* is not the agent's goal over the same interval.
- for any goal an agent has, there is a future interval when the proposition will no longer be the agent's goal; i.e., all goals are eventually dropped.
- goals are believed possible; that is, if an agent has a proposition as a goal, then it believes there is some action it can do which will bring about the desired state of affairs.

Based on the system outlined, we can formulate as shown in Figure 2 the goal of a robot meeting his boss with a cocktail at the train station. This captures the condition that the robot get to the train station before his boss arrives. We wish to emphasize that goals with time constraints cannot even be formulated in C&L's system, unless we assume a clocked world in which all actions take the same amount of time. However, time-constrained goals arise so frequently that we felt ourselves compelled to provide a framework allowing for them.

```
(goal robot '(and (holds (and (loc self station)
                              (holding self cocktail))
                         boss-arrival-time)
                  (occurs (arrives boss) boss-arrival-time))
     now)
```

Figure 2: Example goal

Persistence

Intentions acquire their strength as organizers of our practical action by virtue of their persistence. They would serve no purpose if they were taken up and dropped at whim. Accordingly, an important trait of a goal is how persistent it is. C&L anchor their discussion of intentions around the distinction between persistent goals, which are dropped only when the agent has either achieved them or come to believe them impossible of achievement, and relativized goals. A relativized goal may, for example, be dropped when a plan of which it is part, comes to be seen as infeasible.

The predicate **p-goal** is given in Definition 1. It is a predicate taking the same arguments as **goal**. It makes explicit what will be true when the agent gives up the goal: the agent will either believe the goal holds or that it is impossible. In addition to capturing the essence of C&L's definition, we also require that agents be willing to commit to some plan of action to achieve their goals. Although this is a weak criterion, we consider it important to establish a link between an agent's persistent goals and that agent's future actions.

The "fanatical" commitment captured by **p-goal** is weakened by introducing relativizing conditions, which may be a higher level goal or some belief about the world held by the agent. Relativized goals may be used for purposes like capturing relationships between goals and sub-goals, conditions under which a plan is appropriate and coordinating actions of cooperative agents. Due to space limitations, the definition of **p-r-goal** is omitted.

In order for a reasoning agent to make full use of the system we have presented thus far, they need both positive and negative goal introspection. Consistency of goals requires that agents know what their goals are. Furthermore, agents must know which goals depend on relativizing conditions so that should belief in these conditions cease, the goal can be dropped. Davis[1990, p. 415] provides an axiomatization of goal introspection.

Intended Actions

Bratman describes intentions as corresponding to the mental attitude of having a plan. Thus while the *content* of an intention may be thought of as a plan in the conventional AI sense of a "recipe for action," the intention *per se* is analagous to the "complex mental attitude" described by Pollack [1990].

In Definition 2, we specify the predicate **intend**, which takes as arguments an agent, an action term, an interval within which the action is intended to be done, and an interval over which the agent maintains the intention. **Intend** depends on two predicates the definitions of which are omitted here for space reasons. **Feasible** is true of an action term and an interval when all the preconditions of the action are either true or possible during the interval. **Applicable** is true of an action term, the time of the action's occurrence, a goal, and a bounding interval if the occurrence of the action within the bounding interval either (1) implies the goal proposition, or (2) generates an action the occurrence of which implies the goal state.

Let us now consider the conjuncts on the right hand side of Definition 2. The first two (marked as [a]) allow us to refer to the agent's name for the action and the time within which it is intended to take place. The intention is directed toward some goal ([b]).[2] The state of having the intention is initiated by the event of committing to do the action ([c]). In order for the intention to be rational, we additionally require: [d] that the agent not believe that the action is already done; that the agent believe that the action is possible, and; [f] that the agent believes that the action will serve to achieve a goal.

As **intend** depends on the definition of **p-goal**, we analogously define relative intentions which depend on relativized goals.

Any theory of intention must distinguish between intended and unintended side effects. Bratman argues that the characteristic feature of an unintended side effect is that, should circumstances shift such that the action does not in fact cause the effect, the agent does not view this as a failure and does not replan. Even if, before performing the action, the agent comes to believe that a previously expected side-effect will not occur, he does not replan.

Success and Failure

One use of a theory of intention is that it lends predictability to the actions of rational agents. This in turn gives them a foundation from which to reason about the goals and intentions of other agents. Observations may be made of an agent's actions which allow conclusions regarding that agent's success or failure

[2]Note that this does not preclude actions serving more than one purpose.

Definition 1 (p-goal) *Persistent goals. An agent* x *has a persistent goal* p *at time* t *iff (1)* p *is a goal, (2) the agent believes that eventually he will commit to some action to bring it about that the goal holds, and (3) the agent does not abandon the goal unless he comes to believe either that* p *holds or that it will never hold.*

```
(iff (p-goal x p t)
     (exists s1 s2 m (and (goal x p t)
                          (bel x '(exists ,s1 ,s2 ,m
                                          (and (future ,s2 now)
                                               (occur (commit-to self ,p ,m ,s1) ,s2))) t)
                          (implies (future s t)
                                   (or (p-goal x p s)
                                       (exists s' (and (future s' t)
                                                       (or (bel x p s')
                                                           (bel x '(not (possible ,p)) s'))))))))))
```

Definition 2 (Intend) *Intentions regarding actions, absolute. An action is intended if the agent believes the action feasible and applicable to a p-goal and has committed to that way of making the goal true.*

```
(iff (intend x a at t)
     (exists s1 s2 s3 s4 p m
             (and (= (name-for x a) m) [a]
                  (= (name-for x at) s2)
                  (p-goal x p t) [b]
                  (occur (commit-to x p m s2) s1) [c]
                  (starts s1 t)
                  (bel x '(not (exists ,s3 (and (occur ,m ,s3)
                                                (in ,s3 ,s2)
                                                (< (endpoint ,s3) now))) t)) [d]
                  (bel x '(and (feasible ,m ,s2) [e]
                               (forall ,s4 (applicable ,m ,s4 ,p ,s2))) t)))) [f]
```

and thus guide predictions about what future course the agent is likely to pursue. Such predictions provide a context within which to interpret the agent's future actions.

With respect to predicting success, consider that, by definition, a persistent goal is not dropped unless the agent believes one of two alternatives: that the goal has been met or that it is impossible. By our axiom of limited persistence of goals, we can conclude that an agent will eventually believe one of these alternatives. If it is further assumed that a competent agent never comes to believe that his persistent goal is impossible, then one may conclude the goal is eventually met.

In order to make sense of action sequences, we must recognize when an intention has failed. Informally, we say a failure has occurred when an agent maintains a goal-directed intention after having attempted to bring about his goal state. It follows from the definition of `intend` that an intention will only be given up under one of the following four circumstances:

1. the agent ceases to believe that the action will meet its goal;
2. the agent comes to believe that the action cannot be done (e.g., the agent misses its deadline);
3. the goal is met;
4. the action is performed.

For example, if an agent intends to perform a, performs some action, but does not believe the action he performed was, in fact, a (i.e., he believes he failed to perform a), then the original intention remains. Our formalization thus provides a certain weak, nonmonotonic inference about an agent's willingness to act given his current beliefs and goals. This ability to connect actions to goals allows us to subdivide action sequences according to the goals served. In previous work, we outlined how this interpretation of actions might serve in a formalization of story grammars [Lang and Goldman, 1992]. We are extending that work by using the theory of intention described here as a foundation for formalizing a grammar for narratives. This application is intended both as a step in our ongoing research in NLP and as a way of assessing the representational adequacy of the logical framework.

Discussion

Although we are indebted to C&L for their insights into a formal theory of intention, our theory allows a wider variety of possible goals without sacrificing

C&L's key results. Achieving the same predictions about the persistence of intentions and goals presented difficulties because C&L's proofs relied on the particular restricted class of goals they treat. In order to create a definition of intended action providing useful predictions, we required that there be an *event* in which agents committed to intended actions. We believe that this is not only a theoretical convenience, but also captures an important intuition about prior intentions.

We also feel that basing our theory of intention on a quotational representation of beliefs holds the promise of its extension to more complex models of belief. In particular, unlike modal logics, syntactic theories allow for the possibility of sets of agents whose models of the world differ not only on the properties of objects, but on the domain itself [Maida, 1992].

Allen treats intentional action in [Allen, 1984]. However, Allen's discussion is primarily a demonstration that his logic is expressive enough for discussion of intentions. Rather than a general theory of intentions, he provides examples of the formalization of intentions in particular cases. Little is said about the causal role of intentions in general. Allen discusses an important aspect missing from our theory: the achievement of goals by inaction as well as action. One can often bring about a goal state by failing to do an action one might do "by default." We do not yet know how to analyze this situation.

Our theory captures the intuition that agents are willing to act to bring about their goals. Although we would have liked to say something stronger about an agent's actions based on his goals, we believe we've said as much as possible without introducing the complications counterfactuals would bring into the logic.

We are pursuing applications of this theory to representing the content of narratives and to constructing and understanding description-based communication. These applications require us to consider agents having different world models (separate ontologies). We are also working to incorporate recent refinements of the logic of time into our treatment of intentions [Freksa, 1992, Galton, 1990, Ladkin, 1987].

Acknowledgements We thank Mark S. Boddy of Honeywell SRC, Anthony Maida of University of Southwestern Louisiana and Lynn Andrea Stein of the MIT AI lab for assistance in the research described here.

References

[Allen, 1984] James F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23:123–154, 1984.

[Bratman, 1987] Michael E. Bratman. *Intention, Plans, and Practical Reason.* Harvard University Press, Cambridge, 1987.

[Bratman, 1990] Michael E. Bratman. Intention. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 12–32. MIT Press, 1990.

[Cohen and Levesque, 1990a] Philip R. Cohen and Hector J. Levesque. Intention is choice with committment. *Artificial Intelligence*, 42:213–261, 1990.

[Cohen and Levesque, 1990b] Philip R. Cohen and Hector J. Levesque. Persistence, intention, and commitment. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 33–69. MIT Press, 1990.

[Davis, 1990] Ernest Davis. *Representations of Commonsense Knowledge.* Morgan Kaufmann Publishers, Inc., San Mateo, California, 1990.

[Freksa, 1992] Christian Freksa. Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54(1–2):199–227, 1992.

[Galton, 1990] Antony Galton. A critical examination of allen's theory of action and time. *Artificial Intelligence*, 42:159–188, 1990.

[Haas, 1986] Andrew R. Haas. A syntactic theory of belief and action. *Artificial Intelligence*, 28:245–292, 1986.

[Harel, 1979] D. Harel. *First-Order Dynamic Logic.* Springer-Verlag, New York, 1979.

[Ladkin, 1987] Peter Ladkin. The completeness of a natural system for reasoning with time intervals. In *Proceedings of the IJCAI 1987*, pages 462–467. Morgan Kaufmann Publishers, Inc., 1987.

[Lang and Goldman, 1992] R. Raymond Lang and Robert P. Goldman. Toward a knowledge representation for simple narratives. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, pages 915–920, Bloomington, Indiana, 1992.

[Maida, 1992] Anthony S. Maida. Knowledge representation requirements for description-based communication. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning*, 1992.

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

[Pollack, 1990] Martha E. Pollack. Plans as complex mental attitudes. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 77–104. MIT Press, 1990.