Intentions, Commitments and Rationality

Munindar P. Singh*
Center for Cognitive Science
and Dept of Computer Sciences
University of Texas
Austin, TX 78712
USA

Abstract

Intentions are an important concept in Cognitive Science and Artificial Intelligence (AI). Perhaps the salient property of (future-directed) intentions is that the agents who have them are *committed* to them. If intentions are to be seriously used in Cognitive Science and AI, a rigorous theory of commitment must be developed that relates it to the rationality of limited agents. Unfortunately, the available theory (i.e., the one of Cohen & Levesque) defines commitment in such a manner that the only way in which it can be justified reduces it to vacuity. I present an alternative model in which commitment can be defined so as to have more of the intuitive properties we expect, and be closely connected to agent rationality. This definition is intuitively obvious, does not reduce to vacuity, and has useful consequences, e.g., that a rational agent ought not to be more committed to his means than to his ends.

1 Introduction

Intentions, along with beliefs and desires, are an important component of the folk psychological concepts of intelligence and agency, especially as these concepts are used in Cognitive Science and Artificial Intelligence (AI). Recently, there has been some interest in the formalization of the semantics of intentions—i.e., of the conditions under which an agent may or may not be said to have an intention [Cohen and Levesque, 1990, McDermott, 1982, Singh, 1990, Singh and Asher, 1990].

The modern philosophical view is that intentions cannot be reduced to desires and beliefs (e.g., see [Brand, 1984, pp. 121–125], [Bratman, 1987, pp. 18–23] and [Harman, 1986, pp. 78–79]). Intentions are most often seen as being mutually consistent, compatible with beliefs, and direct or immediate causes of action (e.g., [Brand, 1984, p. 46]). This is a useful property for the purposes of this paper, since it helps relate intentions to rationality via actions. Intentions come in at least two shades: present-directed ones and future-directed ones. It is the latter that will interest us here.

Perhaps the salient property of future-directed intentions is that they involve *commitment* on the part of agents. This view has been gaining ground in the philosophical and AI literatures recently (e.g., see [Bratman, 1987, ch. 2], [Harman, 1986, p. 94] and [Cohen and Levesque, 1990, p. 217]). The idea here is that an agent who has an intention is in some way committed to it—not only does he intend to achieve the relevant condition right now, but would also intend to achieve it later, even as the circumstances changed, perhaps for the worse. Thus there is a certain amount of irrationality built into the very idea of commitment.

Yet there are philosophical as well as practical advantages to the view that agents are (or should be) committed to their intentions. While it admits present-directed intentions (e.g., for actions being done intentionally now), it gives primacy to future-directed ones. This is important since it allows an agent's intentional state *now* to influence his actions later. When conceived of as involving commitments, future-directed intentions allow an agent to coordinate his activities, both with his other activities, and with those of other agents. This is also practically important since it simplifies the design and analysis of complex agents, an important issue in AI.

The commitment-based view of intentions suggests that an agent reconsider his intentions only occasionally, rather than at every step. This allows even a computationally and perceptually limited agent to carry on fairly effectively in a world that, relative to his cognitive and physical capacities, is highly complex and changing rapidly. I take this much as granted in this paper.

In §2, I describe the notion of commitment as applied to intentions in philosophy; in §2.1, I describe how it is formalized in the theory of intentions of Cohen & Levesque [1990]; in §2.2, I point out the major problem with their approach, which seems to trivialize the notion of commitment; and in §2.3, I present my own intuitions about how commitment

can be reconciled with rationality, and propose a definition of it that does not reduce to vacuity. Next, in §3, I explain the ontological framework of this paper. In §5, I present the formal language, and in §6, the formal model. In §7, I return to commitment and try to place it in the context of the formal model presented in §6. The approach presented here is general and independent of the exact semantics given to intentions—be it possible worlds based, sentential or any other.

2 What is Commitment?

An agent's commitment to his intentions differs from his commitment to his beliefs in that only the former can cause the agent to act. Following Bratman and Harman, I consider a mental or internal notion of commitment, rather than a social or external one—an agent is committed (by himself, as it were) to his intentions, not to anyone else. Commitment is thus a purely psychological concept, but has some obvious ramifications on the behavior of agents. As we have seen, it entails that the agent continue to hold on to his intentions over time, even as things get worse. If the circumstances change for the worse, he might try harder, i.e., spend more energy and time on it. E.g., if you are committed to being at the airport at 6:00pm, you would make more than one attempt to hail a taxi; if no taxis are forthcoming you might walk to a better location, rent a car, or request a friend for a ride, and so on.

Commitments help limited agents pursue complex goals that would otherwise be beyond their capacities. Thus, while commitments might prove quite irrational in some cases (e.g., where they lead the agent to do actions that are too expensive, or whose side-effects are too damaging), overall, in at least ordinary circumstances, they are quite rational for agents who cannot think too fast on the fly. E.g., your commitment to be at the airport might make you hijack a bus there (something that you might regret the rest of your life), but such cases of over-commitment are rare (or ought to be rare among rational agents). However, having the commitment saved you from repeatedly planning during the day to be in a neighborhood cafe at 6:00pm.

The moral of this is that (1) if you do not know too much about the present and future state of the world, and have too little time to think, then, on the average, commitments are a good way of being able to get something done; and (2) while you may have commitments, it is not a good idea to over-commit. It is these opposing intuitions about commitment that make it difficult to capture in a reasonably rigorous framework.

2.1 Commitments a la Cohen & Levesque

Cohen & Levesque (hereinafter C & L) agree that commitment is one of the most important characteristics of intentions. They capture commitment as a form of persistence over time in the definition of "persistent goal" [1990, p. 236] and define intentions as special kinds of persistent goals [1990, pp. 245, 248]. This takes care of the positive part of commitments. C & L recognize that this could easily lead to over-commitment and define intentions as a persistent goal that the agent persists with precisely till the point where either (1) he comes to believe that it has been satisfied; or (2) he comes to believe that it will never be satisfied. This is obviously too strong: in many cases an agent should not persist with an intention even though neither of (1) and (2) hold—e.g., Joe can intend to go to Mars, but give up that intention when he realizes that he does not want to suffer through the training. Later in their paper, C & L also allow intentions to be dropped in a third way—when the "reason" for adopting them is no longer valid [1990, pp. 254–255]. However, this would not help in Joe's case: he might still persist with his reason for his original intention, which is to be mentioned in the history books as one of the pioneers of interplanetary travel. Thus it is not easy to give up an intention in this theory.

We can try to weaken C & L's requirement for dropping an intention by (1) generalizing believing that an intention has been achieved to believing that the intended condition would hold even if the agent does not perform any (costly) actions to achieve it; and (2) generalizing believing that an intention is impossible to achieve to believing that it is too expensive to try achieving. Without some motivation on grounds of rationality, C & L would have no legitimate basis for their definitions. Indeed, they are quite explicit that their goal is to capture the *normative* criteria for the "rational balance" between (among other things) (a) committing to, and (b) dropping intentions [1990, p. 214].

2.2 Critique of C & L's Approach

Reasonable though the idea of treating commitment as temporal persistence may seem, it has some major philosophical and technical shortcomings. Remember that all we intuitively wanted as a property of commitment was that it lead to persistence under ordinary circumstances, *not* that it be *identified* with plain persistence. As I described above, C & L's solution to the problem of intentions never being dropped is a special case of the maxim *"intend something as long as it is useful to do so"*; in other words, as long as the intention (or acting for the intention) has a positive expected

utility—the expected utility is negative or zero if the agent believes that the intention has already been achieved or believes that it never will be.

This maxim, which seems to underlie C & L's proposal, is not just true; it is *tautologous!* A rational agent, it says, should have an intention only so long as he believes it to be beneficial to him, all things considered. I.e., at any given time (in any given situation), whether an agent ought to persist with an intention or not depends on whether or not it has positive expected utility for him. But this is really saying that the concept of intentions is redundant in the theory—if the agent is going to look at what is best for him in each situation, then of what use are his intentions to him? And what use is the concept of intentions to us, as theoreticians? In this framework, agents have some memory in that they can continue with their intentions, but must deliberate about these intentions from moment to moment. Thus one of the major philosophical intuitions about the concept of (future-directed commitment-based) intentions is lost (see §1).

In other words, I am arguing that (1) the notion of commitment-based intentions is an important one for Cognitive Science and AI; and (2) C & L's formalization of it (even if weakened and generalized) does not do justice to it. The claim of this paper is that one can, however, understand commitment in a way that can be felicitously formalized, and which avoids the criticism just levied on C & L's theory. In the next subsection, I try to formulate some intuitions about commitment from the point of view of agent rationality.

2.3 Commitment and Rationality: Conative Entrenchment

Intentions are attitudes of rational agents. For agents who are limited, but are rational to some extent, having a commitment is a means of making the effort and time spent on deliberation have a longer term effect than on just the current action—if an agent can commit to an intention or a course of action, he does not have to repeatedly rethink some issues from first principles. By thus committing, the agent would certainly miss out some opportunities that he could have noticed by rethinking, but this comes at the advantage of not having been swamped by deliberation. In many cases, careful deliberation once in a while is better than poor reasoning done repeatedly. And in the long run, the limited agent ought to come out ahead in terms of effort expended and benefits accrued. I take this for granted in this paper.

However, the question I shall address is related to it. Given that commitments are a good idea for the kind of agents and environments that we are considering, one can naturally focus on the normative criteria for determining how committed an agent *should* be to an intention of his. Now the commitment of an agent to an intention is really a measure of the effort he is willing to put in to achieve it, or of the risk he is willing to take in trying to achieve it, or of something along those lines.

Ideally, the commitment of an agent to an intention should depend on its *utility* to him, "utility" here being a normative concept. For a real-life agent, the commitment would actually have to be set equal to the utility he subjectively expects from the intention. This approach has the advantage that once an agent has adopted an intention and decided his level of commitment for it, he does not have to repeatedly reconsider his commitment—he would need to reconsider it only when he had put in effort for it well above his initial commitment, or had tried all the sufficiently low-risk and low-cost means he knows of. At that point he could either drop the intention altogether or reinstate it with a new commitment. Thus, the greater (i.e., larger) the agent's commitment to an intention, the less frequently he would need to reconsider it. To coin a phrase analogous to the well-known for beliefs, an agent's commitment to an intention is a measure of its *conative entrenchment.*

In this paper, I consider only the sense of conative entrenchment in which the expected utility of an intention is involved (rather than risk, or some other such potentially useful criterion). For concreteness, I now turn to a formal model involving action and time in which intentions and commitments can be formalized. This model is quite abstract, is derived from models for branching time temporal logic, has previously been developed, and has been applied to the formalization of intentions and know-how [Singh, 1990, Singh, 1991, Singh and Asher, 1990].

3 The Model, Intuitively

For concepts such as intention, commitment and expected utility to even be formalized, we need a formal model that includes not just time and action, but also possibility, probability and choice. The model I propose here is based on possible worlds. Each possible *world* has a branching *history* of *times.* Histories are sets of times, partially ordered by temporal precedence, $<$. They branch into the future, and are assumed to never end. The sets of times in the history of each world are disjoint. A world and time are a "situation." A *scenario* at a world and time is any maximal set of times containing the given time, and all times that are in a particular future of it; i.e., a scenario is any single branch of the history of the world that begins at the given time, and contains all times in some linear subrelation of $<$. Different scenarios correspond to different ways in which the world may develop as a result of the actions of agents.

Even though a world may develop in several different ways, only one scenario can be actualized. I

take probability as applying to scenarios and denoting their objective chance of being *actualized*. The probability of a scenario is given relative to a world and time and some description of an agent's cognitive state. Even for the same world and time, the description can vary—this allows us to express probability before and after an agent's intentions are considered, and is crucial to the goals of this paper. Only objective probability is considered, but it is seen as being dependent on the agent's internal state, since actions can influence what occurs later, and actions are chosen by agents depending on their internal state. An agent *may* do several basic actions at any world and time; for simplicity, I assume that on each scenario he would do exactly one, intuitively, the one he *chooses* to do.

4 Primitive Concepts

I take *Commits* as a primitive notion here and consider intention as derived. $Commits(x, p, c)$ means that agent x is committed to achieving p to a level of c. Then $Intends(x, p) \equiv (\exists c > 0 : Commits(x, p, c))$. Note that even though commitments can be of different degrees, these degrees just represent the entrenchment of the corresponding intention—an intention itself is treated as being either ON or OFF, i.e., as binary. This is crucial since the motivational component of intentions, which is what makes agents act, is needed fully, if at all, for an agent to act for it—how much effort an agent expends is a different matter (I do not consider actions done half-heartedly: even our pretheoretic intuitions are unclear about such cases).

Each agent deliberates from time to time. $Deliberates(x)$ is true at precisely the situations where x deliberates. The process of deliberation is not studied here, and the theory presented applies only between successive deliberations (on any scenario). Each action when done at a given time along a given scenario has a certain cost attached to it—this cost can vary between different instances of the same action, and equals the value of $Cost(x, a)$ on a given world, time and scenario.

Objective probability is needed in the model to take care of the notion of objective chance. Many actions, e.g., coin tosses or rolls of dice, have several possible outcomes which have (perhaps, different) objective probabilities associated with them. These outcomes may also have different utilities for an agent. In the model, objective probability is treated as a function, $\pi(\cdot)$, from scenarios to the unit interval, $[0 \ldots 1]$ and, utility is captured by a function $\Omega(\cdot, \cdot)$ applied to agents and scenarios. In the language, utility is expressed by a function $Utility(\cdot, \cdot)$ applying to an agent and a condition, and is meant to take the objective chance of different scenarios on which that condition is true into account.

The key feature of intentions that we need is that they lead to action. Intentions here are future-directed and allow not just immediate actions, but also those in the future. Therefore, another useful primitive is *acting for an intention*: an agent acts for an intention when his action is a part of what he would do in order to satisfy it—the details of this process are not focused on here. Acting for an intention is a cognitive concept—it depends on the agent's internal state rather than the world. An agent acting for an intention may be doing so even if it would be impossible or unlikely for him to ever succeed (by doing that action). The same action could be done for two different intentions; of course, several temporally isolated actions may have to be done for a single intention. I notate this concept as a three place predicate $Acts\text{-}for(\cdot, \cdot, \cdot)$: applied to an agent identifier, basic action and a condition. In order to connect the agent's cognitive state with the world, we need the concept of *performing* an action in the world—this is notated by a predicate $Performs(\cdot, \cdot)$. *Performs* of an agent x and action a is true over a subscenario on which x does a. I assume that an agent who acts for a condition intends it, and also immediately performs the action by which he acts for that condition; i.e., $Acts\text{-}for(x, a, p) \rightarrow Performs(x, a)$ is always true.

Commitments (and therefore intentions) and beliefs are given a simple semantics for ease of exposition, and to focus on the matters of interest—a *Commits*, *Believes* or *Acts-for* formula is true over a subscenario or interval if it belongs to the agent's cognitive state during that subscenario. These concepts are treated purely qualitatively; they can straightforwardly by analyzed using subjective probabilities, if one wishes. I consider it a strength of this approach that it does not require the notion of subjective probability, while being compatible with it. Note that agents can have beliefs, and even intentions, that involve objective probability and utility statements.

5 The Formal Language

The formal language of this paper, $\mathcal{L}$, is CTL* (a propositional branching time logic [Emerson, 1989]) augmented with quantification over basic actions; functions: *Prob*, *Utility*, *Cost*; and predicates: *Believes*, *Commits*, *Intends*, *Acts-for* and *Performs*; and the arithmetic required. Let x be an agent; p, q propositions; a an action; and v a probability.

A formula can be any of the following: an atomic formula (ψ), a conjunction of formulae ($p \wedge q$), a negation of a formula ($\neg p$), an until-expression ($p \mathsf{U} q$), an action-expression or any of the four special predicates applied to the appropriate kinds of arguments, or a path-quantifier followed by a formula. A path-quantifier is one of A and E. A denotes "in *all* scenarios at the present time," and E denotes "in *some* scenario at the present time"—i.e., $\mathsf{E}p \equiv \neg\mathsf{A}\neg p$. $\mathsf{F}p$ denotes "p holds sometimes

in the future on this scenario" and abbreviates "**true**Up." G denotes "p always holds in the future on this scenario" and abbreviates "$\neg \mathsf{F} \neg p$." Implication ($p \rightarrow q$) and disjunctions of formulae ($p \vee q$) are defined as the usual abbreviations. An action-expression is of the form $\langle a \rangle p$ and means that action a is done on the given scenario at the given time by agent x, and that p holds as soon as a is done.

6 The Formal Model

The semantics of $\mathcal{L}$ is given relative to intensional models (as described informally in §3): it is standard for CTL*. The formal model is. Let $M = \langle F, N \rangle$ be an intensional model, where $F = \langle \mathbf{W}, \mathbf{T}, <, \mathbf{A}, \mathbf{U} \rangle$ is a frame, and $N = \langle [\![ ]\!], \mathbf{B}, \mathbf{C}, \pi \rangle$ an interpretation. Here $\mathbf{W}$ is a set of possible worlds; $\mathbf{T}$ is a set of possible times ordered by $<$; $\mathbf{A}$ is the class of agents in different possible worlds; $\mathbf{U}$ is the class of basic actions; as described below, $[\![ ]\!]$ assigns intensions to atomic propositions and actions. Scenarios as described in §5 can be defined easily from $<$ [Singh, 1991]; $\mathbf{S}_{w,t}$ is the class of all scenarios at world w and time t: $(w \neq w' \vee t \neq t') \Rightarrow \mathbf{S}_{w,t} \cap \mathbf{S}_{w',t'} = \emptyset$. $\langle S, t, t' \rangle$ is a *subscenario* of S from t to t', inclusive. $\mathbf{B}$ assigns basic actions to the agents at different worlds and times. $\mathbf{C}$ assigns a cognitive state to each agent at different subscenarios, as defined below. A coherence requirement is that the cognitive state for a subscenario cannot be different than for a subscenario containing it. π assigns probabilities to scenarios in $\mathbf{S}_{w,t}$, for each world w and time t.

The intension of an atomic proposition is the set of worlds and times where it is true; that of an action is, for each agent x, the set of subscenarios in the model in which an instance of it is done (from start to finish) by x; e.g., $\langle S, t, t' \rangle \in [\![a]\!]^x$ means that agent x does action a in the subscenario of S from time t to t'. I assume that $[\![ ]\!]$ respects $\mathbf{B}$; i.e., $a \in \mathbf{B}_{w,t}(x)$. For the models to be coherent, we need to constrain them so that (1) an action begun at a time ends at most once on any scenario there; (2) subscenarios are uniquely identified by the times over which they stretch; (3) there is a always future time; and (4) something must be *done* by each agent along each scenario in the model, even if it is a dummy action. Restrictions on $[\![ ]\!]$ can also be used to express the limitations of agents; e.g., x cannot pick up three glasses at once.

The semantics of formulae is given relative to a model as defined above and a world and time in it. $M \models_{w,t} p$ expresses "M satisfies p at w, t." $M \models_{S,t} p$ expresses "M satisfies p at time t on scenario S," and is needed for some formulae as defined in §5. p is *satisfiable* iff for some M, w and t, $M \models_{w,t} p$. p is *valid* in M iff it is satisfiable at all worlds and times in M. The satisfaction conditions for the temporal operators too are adapted from those in [Emerson, 1989]. Formally, we have the following definitions:

1. $M \models_{w,t} \psi$ iff $\langle w, t \rangle \in [\![\psi]\!]$
2. $M \models_{w,t} p \wedge q$ iff $M \models_{w,t} p \wedge M \models_{w,t} q$
3. $M \models_{w,t} \neg p$ iff $M \not\models_{w,t} p$
4. $M \models_{w,t} \mathsf{E}p$ iff $(\exists S : S \in \mathbf{S}_{w,t} \wedge M \models_{S,t} p)$
5. $M \models_{w,t} \mathsf{A}p$ iff $(\forall S : S \in \mathbf{S}_{w,t} \rightarrow M \models_{S,t} p)$
6. $M \models_{S,t} \langle a \rangle p$ iff $(\exists t' : \langle S, t, t' \rangle \in [\![a]\!]^x \wedge M \models_{S,t'} p)$
7. $M \models_{S,t} p \mathsf{U} q$ iff $(\exists t' : M \models_{S,t'} q \wedge (\forall t'' : t \leq t'' \leq t' \rightarrow M \models_{S,t''} p))$

   $p \mathsf{U} q$ is satisfied at time t on scenario S iff there is a time such that q holds at it, and for all times between now and then, p holds at them.
8. $M \models_{S,t} p$ iff $M \models_{w,t} p$, if p is not of the form $q \mathsf{U} r$ or $\langle a \rangle q$, and w is the (unique) world such that $S \in \mathbf{S}_{w,t}$
9. $M \models_{S,t} \mathit{Believes}(x, p)$ iff $(\exists t' : \mathit{Believes}(x, p) \in \mathbf{C}_x(\langle S, t, t' \rangle))$
10. $M \models_{S,t} \mathit{Commits}(x, p, c)$ iff $(\exists t' : \mathit{Commits}(x, p, c) \in \mathbf{C}_x(\langle S, t, t' \rangle))$
11. $M \models_{S,t} \mathit{Acts\text{-}for}(x, a, p)$ iff $(\exists t' : \mathit{Acts\text{-}for}(x, a, p) \in \mathbf{C}_x(\langle S, t, t' \rangle))$
12. $M \models_{S,t} \mathit{Performs}(x, a)$ iff $(\exists t' : \langle S, t, t' \rangle \in [\![a]\!]^x)$
13. $M \models_{w,t} \mathit{Utility}(x, p) = u$ iff

    $$[\sum_{S \in \mathbf{S}_{w,t} \wedge M \models_{S,t} p} \pi_{w,t}(S) \times \Omega_{w,t}(x, S)] = u$$

    The utility of p to x is the weighted sum of the utilities of the scenarios on which p holds.
14. $M \models_{w,t} \mathit{Prob}(p) = v$ iff

    $$[\sum_{S \in \mathbf{S}_{w,t} \wedge M \models_{S,t} p} \pi_{w,t}(S)] = v$$

    The probability of p holding is the sum of the probabilities of the scenarios that satisfy it.

7 Commitment Formalized

Now I return to commitments. An important property of intentions that connects them to action is captured by the following constraint on our models: an agent who has a positive commitment to achieving a condition must eventually act on (unless he deliberates again in the meantime). A stronger formulation would require that an agent did not hold an intention infinitely without acting for it, but there is no space to include such "fairness" conditions here [Emerson, 1989].

1. $\mathsf{A}[\mathit{Intends}(x, p) \rightarrow \mathsf{F}\mathit{Deliberates}(x) \vee \mathsf{F}(\exists a : \mathit{Acts\text{-}for}(x, a, p))]$

Another constraint that we need is the following which essentially "uses up" a part of the commitment to an intention. Here the metaphor of commitment as a measure of the resources committed to an intention is especially attractive. As the agent does actions for his intention, he uses up resources for it, and his intention becomes progressively less entrenched. Finally when his commitment becomes too little, constraint 1 will no longer apply; the agent will no longer be required to act

for that condition (and ordinarily, he would not). He might adopt an intention for the same condition again, in which case he will again be able to do some actions for it.

2. $\mathsf{A}[(\mathit{Commits}(x,p,c) \wedge \mathit{Acts\text{-}for}(x,a,p) \wedge \mathit{Cost}(x,a)=u) \rightarrow \langle a\rangle \mathit{Commits}(x,p,c-u)]$

For contrast, C & L's persistence condition translated to the framework of this paper would be

3. $\mathsf{A}[\mathit{Intends}(x,p) \rightarrow (\mathit{Intends}(x,p)\mathsf{U}(\mathit{Believes}(x,p) \vee \mathit{Believes}(x,\mathsf{AG}\neg p)))]$

This requires that an agent not deliberate about an intention once he adopts it. The proposed framework does not prevent or require such deliberation; e.g., it does *not* require constraint 4, which says that if an agent comes to believe that an intention of his has been satisfied, he has to deliberate immediately.

4. $\mathsf{A}[\mathit{Intends}(x,p) \wedge \mathit{Believes}(x,p) \rightarrow \mathit{Deliberates}(x)]$

Instead, we can have the much weaker constraint 5, which says that when an intention is believed to have succeeded, the agent would eventually deliberate. Essentially the same improvement can be made for the constraint calling for intentions to be dropped when it is believed that the intended is impossible in the future.

5. $\mathsf{A}[\mathit{Intends}(x,p) \wedge \mathit{Believes}(x,p) \rightarrow \mathsf{F}\mathit{Deliberates}(x)]$

Deliberation is also called for when the commitment falls below a certain threshold required to do something.

6. $\mathsf{A}[\mathit{Commits}(x,p,c) \wedge (\forall a : (\mathsf{E}\mathit{Acts\text{-}for}(x,a,p)) \rightarrow \mathit{Cost}(x,a) > c) \rightarrow \mathsf{F}\mathit{Deliberates}(x)]$

This leaves one important requirement still to define. This requirement concerns the definition of commitment as the expected utility of the corresponding intention to the given agent. The following constraint says that if an agent deliberates and adopts an intention, his commitment to that intention equals what he believes is his objectively expected utility of achieving that condition sometimes in the future.

7. $\mathsf{A}[(\mathit{Deliberates}(x) \wedge \mathit{Commits}(x,p,c) \wedge c > 0) \rightarrow \mathit{Believes}(x, \mathit{Utility}(x, \mathsf{F}p) = u)]$

One important aspect of intentions and rationality is that intentions serve as both the *ends* of agents and as their *means* for their other intentions. Joe, in §2.2, has an end of being famous and his means for that end is to go to Mars. Both the end and the means are intentions; however, a rational agent, such as Joe, ought to be more committed to his (ultimate) ends than to the specific means he has chosen—there are usually more than one means to an end, after all. I do not focus here on how an agent might perform the appropriate kind of means-ends reasoning. All that is needed is a formal description of when condition q is seen as a means for p. I propose that q is one of x's means for p iff x intends both and on all scenarios, until he redeliberates all his actions for q are also for p.

8. $\mathit{Means}(x,q,p) \equiv [\mathit{Intends}(x,p) \wedge \mathit{Intends}(x,q) \wedge \mathsf{A}[(\mathit{Acts\text{-}for}(x,a,q) \rightarrow \mathit{Acts\text{-}for}(x,a,p))\mathsf{U}\mathit{Deliberates}(x)]]$

Combining this definition with constraint 2 above, we can easily see that, as desired, the commitment for a means to an end must be less than or equal to the commitment to the end itself.

8 Conclusions and Future Work

I have taken an important property of intentions, commitment, and shown how it can be felicitously reconciled with the idea of rationality for limited agents. It should be clear that limited agents, for whom deliberation is expensive, benefit from being committed to their intentions. Future work planned includes formally deriving that condition, and expressing rationality postulates that relate planning and intentions. Another interesting idea is to define habits as sequences of actions whose cost is lower than the sum of the costs of the individual actions that compose it.

References

[Brand, 1984] Myles Brand. *Intending and Acting.* MIT Press, Cambridge, MA, 1984.

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

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

[Emerson, 1989] E. A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science.* North-Holland Publishing Company, Amsterdam, The Netherlands, 1989.

[Harman, 1986] Gilbert Harman. *Change in View.* MIT Press, Cambridge, MA, 1986.

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

[Singh and Asher, 1990] Munindar P. Singh and Nicholas M. Asher. Towards a formal theory of intentions. In *European Workshop on Logics in Artificial Intelligence*, September 1990.

[Singh, 1990] Munindar P. Singh. Group intentions. In *10th Workshop on Distributed Artificial Intelligence*, October 1990.

[Singh, 1991] Munindar P. Singh. A logic of situated know-how. In *AAAI*, July 1991.