In *Proceedings of the International Joint Conference on Artificial Intelligence*
Workshop on Theoretical and Practical Design of Rational Agents, 1991

On the Commitments and Precommitments of Limited Agents

Munindar P. Singh*

Center for Cognitive Science and Dept of Computer Sciences
University of Texas
Austin, TX 78712
USA

and

DFKI
(German Research Center for AI)
Postfach 2080
D-6750 Kaiserslautern
Germany

msingh@cs.utexas.edu

Abstract

Rationality is an important concept in Artificial Intelligence and Philosophy. When artificial systems are considered to be intelligent or autonomous, it is almost obligatory to attribute intentions and beliefs to them. The currently dominant view of intentions sees them as involving commitments on the part of the agents who have them. But the notion of commitment seems to clash with the notion of rationality. It is argued that this need not be so. Commitments are only appropriate for agents with a limited capacity to reason. A treatment of commitment has been previously proposed that reconciles them with rationality. Here further motivations for the commitments of limited agents are discussed. This analysis is extended to account for the so-called precommitments, which have been excluded by others as introducing too much complexity.

*This research was supported by the National Science Foundation (through grant # IRI-8945845 to the Center for Cognitive Science, University of Texas at Austin) and by the DFKI.

1 Introduction

Rationality is a key concept in AI and related disciplines, including, Economics, Philosophy and Psychology. It provides a powerful abstraction with which to view intelligent agents, independent of their internal architecture and with minimal knowledge about them. Dennett sees the attribution of rationality as the central step in adopting the *intentional stance* [Dennett, 1987]. While agents who are sufficiently autonomous may act in any way they please, with rational agents you have a pretty good idea of how they will proceed, if you know what situation they are in. Rationality constrains the options available to an agent to a small set. This makes it possible to predict with some accuracy, at least in principle, the behavior of rational agents and to ascribe intentions and beliefs to them reliably. If one could design and implement rational agents, one would be a long way along implementing truly intelligent systems. Unfortunately, an assumption usually not noted in the literature on rationality is that real-life agents must necessarily be finite in several respects, among them their ability to compute complex functions under time constraints. Such computations would, in general, be required to take rational decisions.

It is now quite standard in AI and some parts of the related disciplines to note that real-life agents are limited. If rationality is taken as the ability to act effectively despite ones limitations in deciding optimal courses of action, even limited agents can be rational, though not so in the traditional sense. There is now a need for descriptive and prescriptive theories of rationality in limited agents. A *descriptive* theory would consider existing intelligent systems, primarily humans, and describe how they manage to cope despite their limitations; a prescriptive theory would define criteria by which an agent may be designed that exhibits intelligence despite its limitations. This paper considers in a new light some recent descriptive proposals that suggest that humans adopt and hold commitments to the tasks they intend to do. The resulting theory also makes prescriptive sense as an abstract analysis for limited agents.

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 AI. Specifically, intentions and reasoning about intentions are a crucial part of many important subareas of AI—e.g., planning [Georgeff, 1987; McDermott, 1982], plan recognition [Pollack, 1986], natural language understanding [Grosz & Sidner, 1990] and multiagent systems [Singh, 1991c]. 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 [Bratman, 1987, ch. 2] [Harman, 1986, p. 94] [Cohen & 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 some irrationality built into the very idea of commitment. Yet there are philosophical as well as practical advantages to this view. While it admits present-directed intentions,

e.g., for actions being done intentionally now, it gives primacy to future-directed ones. This is important from a philosophical point of view because it allows an agent's intentional state *now* to influence his actions later, in a way that the behavioristically minded philosophers of yore would have found unacceptable [Bratman, 1987, p. 6]. 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 capacities, is highly complex and changing rapidly. I take this much as granted in this paper, and focus on the issue of how the notion of commitment can be understood in AI.

In §2, I describe the notion of commitment as applied to intentions and how it seems to conflict with the notion of agent rationality. In §3, I present my own intuitions about commitment; in §4, I extend them to precommitment. In §5, I explain the ontological framework and primitives used here. In §6, I formalize commitment and in §7 precommitment. The approach presented here is independent of the exact semantics given to intentions—be it possible worlds based, sentential or any other.

2 Commitment

Following Bratman and Harman, I consider a mental notion of commitment, rather than a social one—an agent is committed to his intentions privately, not to anyone else. This kind of psychological commitment is to be distinguished from social commitment (to which it is intimately related, however). Commitment entails that the agent continue to hold on to his intentions over time, even as things get worse. I.e., an agent who is committed to his intentions would try to achieve it again, if his initial attempt was unsuccessful, and possibly try several times. 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.

One of the reasons given to justify the fact that agents are committed to their intentions, and that our theories should admit that agents are so committed, is that 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 fast enough on the fly. E.g., while your commitment to be at the airport might make you hijack a bus there (something that you might regret the rest of your life), such cases of over-commitment

are rare (or ought to be rare among rational agents). However, having that commitment saved you from repeatedly planning during the day to be in a neighborhood cafe at 6:00pm.

The philosophical intuitions here are that (1) if you do not know too much about the 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. Extant theories seem to suggest that an agent ought to be committed to an intention only as long as it is beneficial and ought to give it up as soon as it is not. But then, if the agent has to decide whether a given intention is beneficial or not repeatedly, the concept of commitment is both descriptively and prescriptively redundant—the agent can just do the optimal action at each moment (see [Singh, 1991a] for a detailed discussion). However, commitments are useful when (1) the agent cannot switch tasks quickly; (2) the cost of reasoning is high; (3) the agent cannot consider all relevant aspects of the world on the fly; or (4) the agent has a pretty good model of the world, so that the losses of opportunity are limited.

3 Conative Entrenchment

Intentions are attitudes of rational agents. However, being rational does not entail having unlimited computational power; it just entails being able to use ones resources effectively. 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 intentions to deliberate on. In many cases, careful deliberation once in a while is better than poor reasoning done repeatedly.

A useful consequence of commitments from an AI designer's point of view is that they allow a more modular design than is otherwise possible. The designer has simply to ensure that the agents being designed have the appropriate commitments at certain times or in certain situations. At the next lower level of the design, he must supply a set of means for ensuring that the commitments are met. The interactions between the processes of deliberating about commitments and the processes for acting up to them can thus be streamlined. To a large extent, design of the commitment layer can be carried out independently of the lower layer.

The main advantage of commitments from the agents' point of view is that, at least in the long run, the limited agent ought to come out ahead in terms of effort expended and benefits accrued. This is why agents who have intentions, i.e., deliberate about, adopt, act on, and drop them, do pretty well despite their other limitations. That an agent can actually succeed

in the world has to do with the nature of the world and the design of the agent relative to it. The relevant parts of the world change slowly, and are sufficiently stable so that agents can monitor them in sufficiently large intervals; well-designed agents are able to monitor the relevant parts of the reality that would affect them, and so on.

Given that commitments are a good idea for the kinds 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. Thus it is only reasonable that 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 an advantage in 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 the agent's commitment to an intention, the less frequently he would need to reconsider it. To coin a phrase analogous to the one well-known for beliefs, an agent's commitment to an intention is a measure of its *conative entrenchment*. Note that this account is not entirely accurate for agents who can change their value systems; however, such changes should automatically lead to a reconsideration of all relevant intentions. Here 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).

4 Precommitment

While Bratman presents a commitment-based analysis of intentions, he explicitly rules out cases of what he calls *precommitment*. An agent is precommitted to adopting (or not adopting) an intention if he has decided in advance that he will (or will not) adopt that intention. An agent may adopt a precommitment because he wants to ensure that he will not, in the heat of the moment as it were, make the wrong decision; e.g., an agent may prevent himself from adopting the intention of eating something from his refrigerator by locking it up and then throwing away the key. Bratman seems to think that considerations of precommitment would complicate the relationship between intentions and rationality. Possibly, they would do so slightly.

It seems to me, however, that precommitment is just another example of how a limited agent may try to act rationally. By precommitting to a course of action, the agent makes the results of his careful reasoning carry through longer. An ice-cream addict can save himself a

lot of trouble by making ice-creams inconvenient, or impossible, to obtain. Precommitments of this sort seem to be a canonical way in which limited agents may marshal their resources for deliberation and prevent themselves from being overwhelmed by a complex and rapidly changing world in which their unconsidered actions would usually be suboptimal.

One way in which an agent may adopt a precommitment is by taking out a side bet to do as he *now* thinks is right. While this idea unnecessarily involves the notion of social commitments among agents, it yields the right metaphor with which to think of precommitment. Precommitments just make the corresponding intentions easier or harder to adopt. When, as suggested in §3, commitments are themselves analyzed as the resources allocated to an intention, this makes for a simple treatment of precommitments as well. They may be taken as (1) the amount (positive or negative) that must be added to the utility that would have been computed at run-time to yield the actual commitment, or as (2) the minimum commitment that the agent is allowed to have to the given task then.

While commitment *simpliciter* seems irrational only during the intervals that an agent does not deliberate in, precommitment seems quite blatantly irrational even from the agent's point of view while he deliberates. That is, the agent may know that relative to his beliefs about the utility of the given task what his commitment should be and yet may commit more or less resources to it. The agent appears internally, i.e., even introspectively, inconsistent and irrational. However, this sense of blatant irrationality is tempered by the knowledge that the agent would have about his limitations. If the agent knows he is limited, he might prefer his careful thought to his rushed evaluations, even if the former were based on dated information or on predictions that turned out to be false. Thus precommitments are useful when (1) the agent's tasks are clear cut, so he has to do them anyway; (2) the agent is a poor reasoner under time pressure; or (3) the agent has to commit to other agents about his actions in advance. While commitments hold only up to the next deliberation, precommitments persist through ordinary deliberations and can influence them. I come back to this point in §8.

For concreteness, I now turn to an informal discussion of a formal model in which intentions and commitments can be formalized. This model is quite abstract, is derived from models for dynamic and branching time temporal logics, has previously been developed, and has been applied to the formalization of intentions and know-how [Singh, 1991b; Singh & Asher, 1990]. I follow the presentation of [Singh, 1991a], where it was used to formalize some postulates concerning commitments.

5 Intuitive Description of Model and Primitive Concepts

We need a formal model that involves time, action, possibility and choice and allows some notion of probability so that expected utility may be formalized (see [Singh, 1991a] for details). The model used 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 the 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. The times in the non-overlapping parts of scenarios are incomparable by $<$, but one can assign clock values to them to compare them, and make sense of expressions such as "noon." I shall not include these here for reasons of space. Even though a world may develop in several different ways, only one scenario can be actualized. An agent *may* do any one of several basic actions at any world and time; this along with other agents' actions and events in the environment determine which scenario is actualized.

I take *Commits* as a primitive notion here and consider intention as derived. $\mathit{Commits}(x, p, c)$ means that agent x is committed to achieving p to a level of c. Then $\mathit{Intends}(x, p) \equiv (\exists c > 0 : \mathit{Commits}(x, p, c))$. 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 for them, is needed fully, if at all—how much effort an agent expends for something is a different matter.

Each agent deliberates from time to time. $\mathit{Deliberates}(x)$ is true at precisely the situations where x deliberates. The process of deliberation is not studied here; however, commitments are assigned by it. For commitments, this theory 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 $\mathit{Cost}(x, a)$ on a given world, time and scenario. The function $\mathit{Utility}(\cdot, \cdot)$ applies to an agent and a condition, and takes into account the objective chance of different scenarios on which that condition is true.

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 an intention. 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 distinct and temporally isolated actions may have to be done for a single intention. I notate this concept as a three place predicate $\mathit{Acts\text{-}for}(\cdot, \cdot, \cdot)$: the first argument of which is an agent identifier; the second the basic action done; and the third the condition acted for. In order to connect the agent's cognitive state with the world, 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. Precommitments, notated *Precommitted*, are discussed in §7.

Commitments, precommitments and beliefs are given a simple semantics for ease of exposition, and to focus on the matters of interest—a *Commits*, *Precommitted*, *Believes* or *Acts-for* formula is true over a subscenario or interval if it belongs to the agent's cognitive state during that subscenario. Agents can have beliefs and intentions that involve objective probability and utility statements. In the following, $\mathsf{A}p$ denotes "for all scenarios through this situation p". $\mathsf{E} \equiv \neg\mathsf{A}\neg$. $\mathsf{P}p$ means "p holds sometimes in the past." $\mathsf{F}p$, $p\mathsf{U}q$, $\langle a\rangle p$ all apply on a given scenarios and mean, respectively, that "eventually p holds," "eventually q holds and p holds until then" and "action a is done and p holds as soon as it is done."

6 Commitment Formalized

Several important properties of intentions, e.g., that an agent who has one must eventually act for it unless he redeliberates in the meantime, can be formalized in the framework as presented above. Many of these are given in [Singh, 1991a] and are not repeated here. One of those that we need here is that acting for an intention "uses up" a part of the resources allocated to it. Here the metaphor of commitment as a measure of the resources committed to an intention is invoked. As the agent does actions for his intention, he uses up resources for it, and it becomes progressively less entrenched. Finally when his commitment for the intention is no longer positive, the agent will no longer be required to act for achieving it. He might reinstate that intention, i.e., adopt an intention for the same condition or task again. If he does so, he will again have a positive commitment to it and will be able to do some actions for it.

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

Constraint 2 says that when an intention is believed to have succeeded, the agent would eventually deliberate. This is satisfied if the agent deliberates repeatedly.

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

Constraint 3 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. He does not have to commit to achieving every useful condition.

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

7 Precommitment Formalized

Now I turn to a formalization of the notion of precommitment in the same framework. Let $\mathit{Precommitted}(x,p,c)$ mean that agent x has precommitted to achieving p to the extent of c.

7.1 Precommitment by Deliberative Inertia

This is the kind of precommitment where the agent on adopting a precommitment, simply does not reconsider the corresponding commitment as often as he might have done otherwise; e.g., an agent may vote for his political party, even if he does not rate its candidate very highly. Thus when an agent is precommitted to achieving a certain condition, he would possibly allocate more resources to it than he would have otherwise. The precommitment simply represents the minimum resources that would be assigned to the task. I now redefine the commitments assigned by an agent to an intention to take into account the precommitments he might have. The following definition shows how precommitments can override the current deliberations of an agent.

4. $\mathsf{A}[(\mathit{Deliberates}(x) \wedge \mathit{Believes}(x, \mathit{Utility}(x, \mathsf{F}p) = c) \wedge \mathit{Precommitted}(x, p, d)$
   $\rightarrow \mathit{Commits}(x, p, \max(c, d)))]$

7.2 Precommitment by Elimination of Options

Instead of relying on deliberative inertia, an agent may exhibit his precommitments by simply eliminating certain options, the availability of which might at a later time "tempt" him to consider giving up a commitment too early. An agent may thus "burn his bridges" so to speak and lose the option he would otherwise have of crossing them. In the refrigerator example of §4, the agent exhibits his precommitment, not by decreasing the resources allocated to the relevant intention, but by making the actions available for achieving it more expensive: he would now need to pry open the refrigerator door, or first locate the key. Thus relative to the costs of the task, its utility is modified. Conversely, an agent may do actions that would later make certain intentions more attractive, i.e., increase their utility to him then; e.g., someone may leave his wallet in his office to make sure he returns later to pick it up. Thus he would have to go to his office for his wallet, even if he would not have gone otherwise. This case is considered below. An agent with precommitment d for p does an action of net cost e, but after which his utility for p increases by $d - e$.

5. $\mathit{Precommitted}(x, p, d) \wedge \mathit{Believes}(x, \mathit{Utility}(x, \mathsf{F}p) = c) \rightarrow$
   $(\exists a : \mathit{Believes}(x, \mathit{Cost}(x, a) - \mathit{Utility}(x, \langle a \rangle \neg \mathsf{F}p) = e) \wedge \langle a \rangle \mathsf{true} \wedge$
   $\mathit{Believes}(x, \langle a \rangle \mathit{Utility}(x, \mathsf{F}p) = c + d - e))$

8 Conditional Commitment and Conative Policies

Another form of commitment that is relevant to limited rational agents may be dubbed *conditional commitment*. This is a commitment that an agent would have for an intention or

task, were a certain condition to obtain. It should be noted that a conditional commitment in q relative to p is different from a commitment in the conditional expression $p \rightarrow q$. The former leads to a commitment on part of the agent (for q) only when the antecedent, p, becomes true; the latter holds anyway. The former may only be satisfied by achieving q; the latter may also be satisfied by achieving $\neg p$, i.e., by making $p \rightarrow q$ vacuously true. Also, the former is really a nested commitment, i.e., a commitment to have a certain commitment. It is easy to see that conditional commitment generalizes the notion of precommitment.

Conditional commitments control the commitments that an agent can come to adopt when he deliberates. They can thus be seen as embodying certain kinds of *conative policies* that an agent might have. Conative policies, akin to their well-known cousin, epistemic policies, are about the kinds of intentions an agent would adopt under various circumstances. One can impose constrains on the conative policies of agents, e.g., to prevent them from adopting intentions which they believe are mutually inconsistent or inconsistent with their beliefs. Conative policies are best framed as different kinds of rationality postulates on an agent's intentions. The conative policies embodied in an agent do not change due to ordinary deliberations. Deliberations of a deeper nature, on par with value assignments, are needed to create and modify them.

9 Conclusions and Future Work

Commitments and precommitments are important aspects of rational agency in human beings. The theory presented here is sufficiently abstract to capture our essential intuitions about these concepts. It is suggested that commitments and precommitments are important components of any descriptive theory of rationality in limited agents, such as humans. This motivates their use in prescriptive theories for AI agents, whose resource limitations cannot be neglected. In particular, it is proposed that commitments be analyzed as the resources that an agent ought to allocate to different tasks. This leads to some interesting properties of commitments, and turns out to be powerful enough to formalize precommitments with. Thus progress is made towards a clearer understanding the rationality of limited agents.

Future work planned includes formally expressing rationality postulates that relate planning and intentions and to consider nested applications of it, as required for plan recognition. These would be done by considering several kinds of interesting conative policies. 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 them.

References

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

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

[Dennett, 1987] Dennett, Daniel C.; 1987. *The Intentional Stance.* MIT Press, Cambridge, MA.

[Georgeff, 1987] Georgeff, Michael P.; 1987. Planning. In Traub, J. F., editor, *Annual Review of Computer Science, Vol 2.* Annual Reviews Inc., Palo Alto, CA.

[Grosz & Sidner, 1990] Grosz, Barbara and Sidner, Candace; 1990. Plans for discourse. In Cohen, P.; Morgan, J.; and Pollack, M., editors, *SDF Benchmark Series: Intentions in Communication.* MIT Press, Cambridge, MA.

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

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

[Pollack, 1986] Pollack, Martha E.; 1986. *Inferring Domain Plans in Question Answering.* Ph.D. Dissertation, University of Pennsylvania.

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

[Singh, 1991a] Singh, Munindar P.; 1991a. Intentions, commitments and rationality. In *13th Annual Conference of the Cognitive Science Society.*

[Singh, 1991b] Singh, Munindar P.; 1991b. A logic of situated know-how. In *National Conference on Artificial Intelligence (AAAI).*

[Singh, 1991c] Singh, Munindar P.; 1991c. Towards a formal theory of communication for multiagent systems. In *International Joint Conference on Artificial Intelligence (IJCAI).*