Commitments in the Architecture of a Limited, Rational Agent* **

Munindar P. Singh***

Department of Computer Science
North Carolina State University
Raleigh, NC 27695-8206, USA

singh@ncsu.edu

Abstract. Rationality is a useful metaphor for understanding autonomous, intelligent agents. A persuasive view of intelligent agents uses cognitive primitives such as intentions and beliefs to describe, explain, and specify their behavior. These primitives are often associated with a notion of *commitment* that is internal to the given agent. However, at first sight, there is a tension between commitments and rationality. We show how the two concepts can be reconciled for the important and interesting case of limited, intelligent agents. We show how our approach extends to handle more subtle issues such as *precommitments*, which have previously been assumed to be conceptually too complex. We close with a proposal to develop *conative policies* as a means to represent commitments in a generic, declarative manner.

1 Introduction

How can limited agents cope with a complex world? This is a question that has drawn much attention in the study of intelligent agents. As agents find application in an increasing variety of complex domains, this question continues to gain importance. There are two dominant views about intelligent agency.

- *Cognitive:* The cognitive view borrows folk psychological metaphors, and treats agents as loci of beliefs, desires, intentions, and so on. This view is called the *knowledge level* [Newell, 1982] or the *intentional stance* [McCarthy, 1979].
- *Economic:* The economic view borrows economic metaphors, and treats agents as rational beings. It has long been realized that perfect rationality is not realizable in limited agents, and theories of bounded rationality have been proposed [Simon, 1981].

* This paper synthesizes and enhances some ideas that were introduced in papers that appear in the Proceedings of the 13th Annual Conference of the Cognitive Science Society (1991) and the Proceedings of the IJCAI-91 Workshop on the Theoretical and Practical Design of Rational Agents.

** I am greatly indebted to Lawrence Cavedon for extensive comments.

*** This work has been partially supported by the NCSU College of Engineering and by the National Science Foundation under grants IRI-9529179 and IRI-9624425.

(Yet another view treats intelligence as emerging from the reactive behaviors of agents. We shall discuss this view further below.) A number of agent architectures based on the above views have been proposed. We propose not a new architecture, but a knowledge-level characterization of an architecture for limited, rational agents. Our goal is to relate the above cognitive and economic metaphors, and to formalize some primitives through which limited, rational agents can be described at the knowledge level. This is a challenging goal because, as we show below, the two views appear incompatible at first glance. The primitives we propose are related to the *intentions* of an agent. One primitive is *commitment*, which is well-known in the literature; another primitive is *precommitment*, which has been mentioned but largely ignored in the literature. We formalize these primitives in a manner that satisfies the intuitions of both the cognitive and the economic views.

There is need for both 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. Such a theory might also be used by artificial agents to understand each other. A *prescriptive* theory would define criteria by which an agent may be designed that exhibits intelligence despite its limitations. This paper first considers commitments in a new descriptive light. From this analysis, it motivates a prescriptive theory that applies to artificial, limited agents.

Section 2 describes commitments as they apply to intentions. It discusses the relationship between commitments and rationality. Section 3 conceptually describes our approach, which is based on treating commitments as levels of entrenchment. Section 4 identifies and describes some key concepts that underlie a formalization of commitments. Section 5 applies the above concepts to formalize and evaluate various constraints about entrenchment, culminating in a call for general conative polices. The formal semantics is outlined in the appendix.

2 Commitments and Rationality

The cognitive view of agency leads to a BDI architecture—one which assigns beliefs, desires, and intentions to agents [Rao & Georgeff, 1991]. Beliefs describe the information available to an agent; desires describe an agent's wants; intentions describe what an agent wants and has decided to act upon. Our interest here is in intentions. Intentions denote an agent's pro-attitude toward a proposition or action. Intentions are usually defined to be mutually consistent, compatible with beliefs, and direct or immediate causes of action (e.g., [Brand, 1984, p. 46]). For the above reasons, intentions are distinct from desires (which may be mutually inconsistent or incompatible with beliefs, and may not lead to actions) and beliefs (which do not in themselves lead to action). This view is supported by a number of philosophers, e.g., [Brand, 1984, pp. 121–125], [Bratman, 1987, pp. 18–23], and [Harman, 1986, pp. 78–79]. We restrict ourselves to intentions that are *future-directed*, i.e., geared toward future actions or conditions.

The literature over the past decade or so agrees on the idea that intentions

involve some *commitment* on part of the given agent. This commitment is "psychological" rather than "social" [Castelfranchi, 1995; Singh, 1996]. An agent is committed privately to his intentions, independently of his public obligations.

2.1 Why Commitments are Useful

Commitments cause an agent to continue to hold on to his intentions over time, and to try repeatedly to achieve them.

Example 1. Consider being committed to going to the airport at 6:00 PM. Then, 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, request a ride, or find some other means to get to the airport on time. ∎

From an agent's standpoint, a useful consequence of commitments is that they enable his intentional state *now* to influence his actions later. Commitments enable an agent to coordinate his activities, both with his other activities, and with those of other agents.

Example 2. Having a commitment to go to the airport will save you the trouble of repeatedly planning to go to a university cafe at 6:00 PM, which you wouldn't be able to act on if you keep your original commitment. ∎

From an agent designer's standpoint, a useful consequence of commitments is that they enable 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, the designer must supply a set of means, e.g., a plan library, 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, the design of the commitment layer can be carried out independently of the lower layer.

Example 3. High-level considerations, e.g., communications or social norms, may lead to the adoption of a commitment, but it is up to the planning and execution module to effect the necessary actions. You can commit to going to the airport based on a phone call from a friend, but whether you get there may depend on your driving, rather than your linguistic, skills. ∎

2.2 How Commitments can be Harmful

Commitments can be harmful when they cause agents to behave suboptimally or irrationally. Commitments essentially work by taking the decision away from an agent in a specific situation, by making the agent act based on his prior judgment and his dated knowledge. If treated qualitatively, commitments can lead to actions that are unduly expensive.

Example 4. Your commitment to be at the airport might make you go there even though your trip was canceled. ▮

Example 5. Your commitment to be at the airport might make you hijack a bus (something that you might regret the rest of your life). ▮

Commitments must be tempered in some way to avoid situations where an agent latches on to his commitment fanatically.

2.3 Commitments versus Rationality, Descriptively

Commitments help limited agents pursue complex goals that would otherwise be beyond their capacities. Thus, while commitments might prove quite irrational in some cases, overall, at least in ordinary circumstances, they are quite rational for agents who cannot think too fast on the fly. This requires that over-commitment ought to be rare, or at least be bounded. Conceptually, we have that

- if an agent has partial knowledge about the future state of the world, and has too little time to think, then, on the average, commitments are a good way of being able to get something done
- it is not a good idea to over-commit.

Commitments pay off in the long run, because cognitive agents can manage to commit without over-committing. This is of course a matter of good design—agents should match the world they exist in. The relevant parts of the world are stable enough that agents can monitor them in sufficiently large intervals.

Example 6. A trip would typically be canceled sufficiently in advance or be a significant enough event that you will end up deliberating (and giving up the commitment to get to the airport) before expending too much effort. ▮

Intuitively, 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.

2.4 Traditional Approaches

Briefly, traditional theories, e.g., [Rao & Georgeff, 1991; Cohen & Levesque, 1990], appear 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. However, 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 perform the optimal action at each moment! Indeed, this unwittingly supports the position taken by [Brooks, 1991] and others that cognitive concepts can be dispensed with entirely in the study of agents. Our chief reason for including cognitive concepts, however, is that they provide a high-level, flexible, declarative means to describe agents.

Commitments as Persistence A well-known traditional approach captures commitment as a form of persistence over time (this is in the definition of "persistent goal") [Cohen & Levesque, 1990, p. 236]. Intentions are defined as special kinds of persistent goals (pp. 245, 248, 254–255). An intention is defined as a goal that the agent persists with precisely until he comes to believe that

PERS-1. it has been satisfied;
PERS-2. it will never be satisfied; or
PERS-3. the "reason" for adopting it is no longer valid.

This characterization is obviously too strong: in many cases an agent should not persist with an intention even though the above do not hold.

Example 7. Joe intended to go to Mars. He would like to give up that intention when he realizes that he does not want to suffer through the training. By the above theory, he cannot! Note that PERS-3 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. ∎

Clearly, it is not easy to give up an intention in this theory. We can try to weaken the above requirements for dropping an intention by replacing the set PERS-1, PERS-2, and PERS-3 by the set PERS-4 and PERS-5. That is, the agent persists with an intention until he comes to believe that

PERS-4. success is inevitable, i.e., the intended condition would hold even if the agent does not perform any (costly) actions to achieve it; or
PERS-5. success is unaffordable, i.e., achieving the intention is too expensive.

Example 8. Continuing with Example 7, Joe can now give up his intention to go to Mars if he believes that the glory is not worth the pain. ∎

Problems with Persistence Reasonable though the idea of treating commitment as temporal persistence may seem, it has conceptual and practical shortcomings. Even the weakened postulates PERS-4 and PERS-5 are a special case of the maxim *"intend something as long as it is useful to do so"*. In other words, an intention is held as long as it has a positive expected utility—the expected utility is negative or zero when PERS-4 and PERS-5 are satisfied. This maxim is eminently rational. It says that a rational agent should hold an intention only so long as he believes the intention to be beneficial, all things considered.

However, in taking care of rationality, this maxim makes the cognitive concepts theoretically redundant. This is because it requires an agent to engage in deep reasoning about his intentions at each moment. If the agent can perform such reasoning effectively, he might just decide what actions are optimal, and save all the bother of having intentions. Thus one of the major conceptual intuitions about intentions and commitments is lost.

Intuitively, commitments should ordinarily *lead to* persistence. The traditional approaches err in *identifying* commitments with persistence.

3 Entrenchment

The essence of commitments is in avoiding having to repeatedly reason about one's actions. We call this *entrenchment*. Persistence is a natural consequence of entrenchment. *Conative entrenchment* applies to the entrenchment corresponding to a simple commitment to an intention. *Deliberative entrenchment* applies to the entrenchment corresponding to a *precommitment*, which we introduce below.

3.1 Conative Entrenchment

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. A committed agent would certainly miss out some opportunities that he could have noticed by rethinking, but at the advantage of not being swamped by intentions to deliberate on. 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) for having committed. Let us restrict ourselves to environments where this is the case.

We propose that the commitment of an agent to an intention is a measure of the time or effort he is willing to put in to achieve it, or of the risk he is willing to take in trying to achieve it. Once an agent has adopted an intention, and decided his level of commitment for it, he would need to reconsider it only when his time or effort or risk exceeds his initial commitment. There is obviously some computation required to keep track of when to reconsider, but in our approach it is relatively small. At that point he could either drop the intention altogether or reinstate it with a fresh commitment. Thus, the greater the agent's commitment to an intention, the less frequently he would need to reconsider it.

Given that commitments correspond to entrenchment, the next natural question is *are there any normative or prescriptive criteria for determining how much an agent should be committed to an intention?* We propose that there can be several normative criteria depending on how one chooses to design an agent. Two of the possible candidates are introduced below (other variations are of course possible). This paper concentrates on the first candidate.

Utilitarian Entrenchment This approach seeks to maximize expected utility. The commitment of an agent to an intention depends on the utility of that intention. 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 limits the effort invested by an agent in satisfying a commitment. An important special case is when the cost is set to the time taken to perform an action.

Entrenchment Through Risk Aversion This approach seeks to minimize the total risk that an agent will undertake in order to satisfy a commitment. Thus it eliminates actions that are highly risky.

3.2 Deliberative Entrenchment

Although Bratman presents a commitment-based analysis of intentions, he explicitly rules out cases of *precommitment* (p. 12). An agent is precommitted to adopting (respectively, not adopting) an intention if he has decided in advance that he will (respectively, will not) adopt that intention. An agent may precommit because he wants to ensure that he will not, in the heat of the moment as it were, make the wrong decision.

Example 9. An agent may prevent himself from adopting the intention of eating ice-cream from his refrigerator by locking it up, and throwing away the key. ∎

Bratman, however, rules out precommitments, because they would complicate the relationship between intentions and rationality. It appears, however, that precommitments are just another example of how a limited agent may try to act rationally. In our formalization, the complexity they introduce is minimal. 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 enable limited agents to marshal their resources for deliberation, and avoid being overwhelmed by a complex world in which their unconsidered actions would likely be suboptimal.

While commitments can cause irrationality only between successive deliberations, precommitments are quite blatantly irrational even while deliberating. 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 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. We conjecture that 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.

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 precommitments.

Example 10. Intuitively, the would-be dieter can make a side bet against his eating the ice-cream, making the cost of having the ice-cream greater than the benefit. This is one of the forms of precommitment that we formalize below. ∎

Precommitments make the associated commitments more or less entrenched, or the corresponding intentions easier or harder to adopt. When 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

- bounds on the commitment that the agent would assign to an intention; or
- the amount (positive or negative) that must be added to the utility that would have been computed to yield the actual commitment.

4 Primitive Concepts

We now introduce some primitive concepts through which commitments and precommitments can be formalized. These concepts involve probabilistic and utilitarian generalizations of a framework previously used to give a logical characterization of intentions and know-how [Singh, 1994; Singh & Asher, 1993].

4.1 Actions, Branching Time, Probabilities

For concepts such as intentions, commitments, and expected utility to be formalized, we need a model that includes not just time and action, but also possibility, probability, and choice. Our model is based on a set of possible moments, partially ordered by time. The moments represent possible states or snapshots of the world. A partial ordering means that a number of *scenarios* may branch out into the future of each moment, each scenario representing a different course of events along which the world may evolve. At each moment, each agent can choose from a number of action instances, one on any scenario. Of the scenarios, only one may be *realized*. Our model assigns probabilities to the scenarios denoting their objective chance of being realized.

4.2 Cognitive Primitives

We take commitments as primitive, and intentions as derived. $\mathsf{C}_x(p, c)$ means that agent x is committed to achieving p to a level of c. Then $\mathsf{I}_x(p) \equiv (\exists c > 0 : \mathsf{C}_x(p, c))$. $\mathsf{I}_x(p)$ means that agent x intends p. For simplicity, unlike [Singh, 1994], we assume that p has an explicit temporal component to capture the future-directedness of intentions. 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. Precommitments are notated by PreC—$\mathsf{PreC}_x(p, c)$ means that agent x has precommitted to achieving p to the extent of c. $\mathsf{Delib}(x)$ is true precisely when agent x deliberates.

Utility is expressed by a function $\mathsf{Uti}(\cdot, \cdot)$ applying to an agent and a condition, which is evaluated at a given moment. Utilities are based on the agents' value systems, and are given as primitives. Each action instance has a cost—this cost can vary across different instances of the same action, and is given as $\mathsf{Cost}_x(a)$ on a given moment and scenario. Many actions, e.g., coin tosses or rolls of dice, have several possible outcomes which have (perhaps, different) objective

probabilities associated with them. Objective probability is given by a function, $\Pi(\cdot)$, from scenarios to the unit interval $[0 \ldots 1]$. The expected cost of an action can be computed based on the probabilities of the different scenarios along which the action may progress, and its cost along each scenario.

A key feature of intentions is that they lead to action. 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. An agent acting for an intention may be doing so even if it would be impossible or unlikely for him to ever succeed though that action. The same action could be performed for two different intentions; of course, several temporally isolated actions may have to be performed for a single intention. Acting for is notated by For. $\mathsf{For}_x(a, p)$ means that agent x performs action a for condition p. Agents can have beliefs and intentions that involve objective probability and utility statements.

A formal semantics is presented in Appendix A.

4.3 Formal Language

The formal language of this paper, $\mathcal{C}$ (for *CONATE*), is based on CTL* (a propositional branching time logic [Emerson, 1990]). It is augmented with (1) quantification over basic actions; (2) functions: Uti, Cost; (3) operators: B and $\langle\,\rangle$; (4) predicates: C, PreC, Delib, I, and For; and (5) arithmetical operators and relations. Let x be an agent; p, q propositions; a an action; and v a probability.

Now we define the syntax of $\mathcal{C}$. We assume that Φ is a set of atomic formulae; $\mathcal{A}$ is a set of action symbols; $\mathcal{X}$ is a set of agents; and $\mathcal{T}$ is a set of terms. $\mathbb{R}$ is the set of reals: $\mathbb{R} \subseteq \mathcal{T}$. $\mathcal{C}$ may be defined by the following rules, which simplify the syntax of CTL* for ease of exposition.

C1. Atomic formulae: $\phi \in \mathcal{C}$, for all $\phi \in \Phi$
C2. Conjunction: $p, q \in \mathcal{C} \Rightarrow p \wedge q \in \mathcal{C}$
C3. Negation: $p \in \mathcal{C} \Rightarrow \neg p \in \mathcal{C}$
C4. Action: $a \in \mathcal{A}, p \in \mathcal{C}, x \in \mathcal{X} \Rightarrow \langle a \rangle_x p \in \mathcal{C}$
C5. Until: $p, q \in \mathcal{C} \Rightarrow p \mathsf{U} q \in \mathcal{C}$
C6. Scenario-quantifier: $p \in \mathcal{C} \Rightarrow \mathsf{A} p \in \mathcal{C}$
C7. Action-quantifier: $p \in \mathcal{C}$, $a \in \mathcal{A} \Rightarrow (\bigvee a : p) \in \mathcal{C}$
C8. Belief: $p \in \mathcal{C}$, $x \in \mathcal{X} \Rightarrow \mathsf{B}_x p \in \mathcal{C}$
C9. Commitment: $p \in \mathcal{C}$, $c \in \mathcal{T}$, $x \in \mathcal{X} \Rightarrow \mathsf{C}_x(p, c) \in \mathcal{C}$
C10. Precommitment: $p \in \mathcal{C}$, $c \in \mathcal{T}$, $x \in \mathcal{X} \Rightarrow \mathsf{PreC}_x(p, c) \in \mathcal{C}$
C11. Acts-for: $p \in \mathcal{C}$, $x \in \mathcal{X}$, $a \in \mathcal{A} \Rightarrow \mathsf{For}_x(a, p) \in \mathcal{C}$
C12. Utility: $p \in \mathcal{C}$, $r \in \mathbb{R} \Rightarrow (\mathsf{Uti}(p) = r) \in \mathcal{C}$

The operators $\wedge$ and $\neg$ are the classical boolean operators. Implication ($p \rightarrow q$) and disjunctions of formulae ($p \vee q$) are defined as the usual abbreviations. true abbreviates $p \vee \neg p$, for any atomic proposition p. false abbreviates $\neg\mathsf{true}$.

An action-expression of the form $\langle a \rangle_x p$ means that action a is performed at the given moment along the given scenario by agent x, and that p holds as soon

as a is completed. $[a]_x p$ abbreviates "$\neg\langle a\rangle_x \neg p$" and means that if a is performed, then p holds upon its completion. "x" is omitted when understood.

$\bigvee$ and $\bigwedge$ are restricted quantifiers that apply only to actions. $(\bigvee a : p)$ means that there is an action which when substituted for a in p yields true—thus it corresponds to an existential quantifier. $(\bigwedge a : p)$ abbreviates "$\neg(\bigvee a : \neg p)$" and corresponds to a universal quantifier.

$p\mathsf{U}q$ is satisfied at moment t on a scenario iff q holds on a future moment on that scenario and p holds at each moment between now and the given occurrence of q. $p\mathsf{U}q$ entails that q holds eventually. $\mathsf{F}p$ denotes "p holds sometimes in the future on this scenario" and abbreviates "$\mathsf{true}\mathsf{U}p$." $\mathsf{G}p$ denotes "p always holds in the future on this scenario" and abbreviates "$\neg\mathsf{F}\neg p$." A scenario-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"—that is, $\mathsf{E}p \equiv \neg\mathsf{A}\neg p$.

Example 11. (1) $\langle a\rangle_x \mathsf{true}$ means that x performs a at the given moment on the given scenario. (2) $\langle a\rangle \mathsf{C}_x(p, c-u)$ means that immediately after a, x's commitment to p is $c-u$. (3) $\mathsf{B}_x(\mathsf{Uti}_x(p) - \mathsf{Cost}_x(a) = e)$ means that x believes that, for him, the difference in the utility of achieving p and the cost of performing a is e. (4) $(\bigvee a : \mathsf{For}_x(a,p))$ means that x performs some action for p. (5) $\mathsf{F}(\bigvee a : \mathsf{For}_x(a,p))$ means that eventually x will perform some action for p. (6) $\mathsf{AG}p$ means that p will hold at all moments on any (future) scenario. ∎

$\langle a\rangle_x \mathsf{true}$ captures the notion of agent x (currently) performing action a. We require that an agent who acts for a condition intends it, and immediately performs the corresponding action. That is, $\mathsf{For}_x(a,p) \rightarrow \langle a\rangle_x \mathsf{true}$ always holds.

5 Entrenchment Formalized

5.1 Conative Entrenchment

We consider only utilitarian entrenchment below. An important property of intentions that connects them to action is captured by the following constraint: an agent who has a positive commitment to achieving a condition must eventually act on it (unless he deliberates again in the meantime). The following constraint says that at all scenarios from any moment, if x intends p then eventually x will deliberate or eventually x will act on p. We omit the agent subscript throughout this section, because the constraints all involve only agent x.

D1. $\mathsf{A}[\mathsf{I}(p) \rightarrow \mathsf{F}(\bigvee a : \langle a\rangle \mathsf{Delib}(x) \vee \mathsf{For}(a,p))]$

The following constraint essentially "uses up" an agent's commitment to an intention. As the agent acts for his intention, his intention becomes progressively less entrenched. Finally when his commitment for the intention is no longer positive, the constraint for motivation (D1 above) will no longer apply; so the agent will no longer be required to act for achieving that condition. Under ordinary circumstances, the agent will no longer act for that intention. He might reinstate that intention (i.e., adopt an intention for the same condition again), in which case he will again be able to act for it.

D2. $\mathsf{A}[(\mathsf{C}(p,c) \wedge \mathsf{For}(a,p) \wedge \mathsf{Cost}(a) = d) \rightarrow \langle a \rangle \mathsf{C}(p, c-d)]$

For contrast, PERS-1 and PERS-2 both translate into constraint D3 in our framework. Intuitively, constraint D3 says that if x intends p then either he intends p forever or intends p until he comes to believe p or believe that p is impossible. The traditional theories separately require that all goals and intentions are eventually dropped. This would eliminate the $\mathsf{GI}(p)$ subexpression below, but causes other repercussions, which are discussed in [Singh, 1992].

D3. $\mathsf{A}[\mathsf{I}(p) \rightarrow (\mathsf{GI}(p) \vee (\mathsf{I}(p) \mathsf{U} (\mathsf{B}(p) \vee \mathsf{B}(\mathsf{AG} \neg p))))]$

The proposed framework does not require constraint D4 either, which says that if an agent has a certain intention, and comes to believe that it has been satisfied, then he has to deliberate immediately.

D4. $\mathsf{A}[(\mathsf{I}(p) \wedge \mathsf{B}(p)) \rightarrow \mathsf{Delib}(x)]$

Instead, we can have the weaker constraint D5, which says that when an intention is believed to have succeeded, the agent will eventually deliberate.

D5. $\mathsf{A}[(\mathsf{I}(p) \wedge \mathsf{B}(p)) \rightarrow \mathsf{FDelib}(x)]$

Essentially the same improvement can be made for the constraint calling for intentions to be dropped when it is believed that the intended condition is impossible along any scenario.

D6. $\mathsf{A}[(\mathsf{I}(p) \wedge \mathsf{B}(\mathsf{AG} \neg p)) \rightarrow \mathsf{FDelib}(x)]$

This leaves one important requirement, which concerns the definition of commitment as the utility of the corresponding intention. The following constraint says that if an agent deliberates, and adopts an intention, his commitment to that intention equals his believed utility of achieving that condition. He does not have to commit to achieving every useful condition.

D7. $\mathsf{A}[(\mathsf{Delib}(x) \wedge \mathsf{C}(p,c) \wedge c > 0) \rightarrow \mathsf{B}(\mathsf{Uti}(p) = c)]$

Constraints Putting together the above, we require constraint D1, D5, D6, and D7. The above constraints handle cases of irrational over-commitment. However, additional primitives, e.g., those of [Norman & Long, 1996], are needed to cause agents to deliberate when the reasons for an intention are invalidated. We lack the space to formalize these here.

Means and Ends Rational agents must relate their means to their ends. Intentions correspond to ends, and plans to means. The plans can lead to additional intentions, which apply within the scope of the original intention. We discuss this issue briefly to outline some of the considerations a formalization of commitments would need to accommodate.

Example 12. Continuing with Example 7 of section 2.4, let us assume that Joe has an end of being famous, and his means for that end is to go to Mars. ▌

We propose the following maxim: *a rational agent ought to be more committed to his (ultimate) ends than to any specific means.* This maxim is reasonable, because there are usually more than one means to an end, after all. We define $\mathsf{Best}(p)$ as the best plan for achieving p. (We have not formalized plans, but one can take them below as partial orders of actions.) For a plan ξ, condition p, and scalar c, we define $\mathsf{Yield}(\xi, p, s)$ to mean that the plan ξ yields a subsidiary commitment of entrenchment s toward p. Then the above maxim translates into the following formal constraint.

D8. $\mathsf{A}[(\mathsf{Delib}(x) \wedge \mathsf{C}(p, c) \wedge c > 0) \rightarrow (\sum_{\mathsf{Yield}(\mathsf{Best}(p), q, s)} s \leq c)$

The above constraint states that the total of the commitments to the subsidiary intentions are bounded by the commitment to the original intention. Greater sophistication is required before the subtleties of the interplay between rationality and intentions can be fully captured.

5.2 Deliberative Entrenchment

Now we turn to a formalization of the notion of precommitment in the above framework. We now redefine the commitments assigned by an agent to an intention to take into account the precommitments he might have. The following constraints shows how precommitments can override current deliberations.

Precommitment by Deliberative Inertia By deliberative inertia, we mean that the agent, on adopting a precommitment, simply does not reconsider the corresponding commitment as often as he might have otherwise.

Example 13. An agent may continue to have ice-cream out of habit, even though he would not do so were he to examine his diet carefully. ▌

By conative entrenchment, an agent's commitment to an intention would peter out in due course. Deliberative inertia makes it last longer. Thus when an agent is precommitted to achieving a certain condition, he would possibly allocate more resources to it than he would have otherwise.

Lower Bound Hysteresis. An agent might precommit by setting the minimum resources that would be assigned to the given commitment.

D9. $\mathsf{A}[(\mathsf{Delib}(x) \wedge \mathsf{B}(\mathsf{Uti}(p) = c) \wedge \mathsf{PreC}(p, d)) \rightarrow \mathsf{C}(p, \max(c, d))]$

Upper Bound Hysteresis. An agent might precommit by setting the maximum resources that would be assigned to the given commitment.

D10. $\mathsf{A}[(\mathsf{Delib}(x) \wedge \mathsf{B}(\mathsf{Uti}(p) = c) \wedge \mathsf{PreC}(p, d)) \rightarrow \mathsf{C}(p, \min(c, d))]$

Additive Bias. An agent might precommit by adding (positive or negative) resources to a commitment when the conative entrenchment is computed.

D11. $\mathsf{A}[(\mathsf{Delib}(x) \wedge \mathsf{B}(\mathsf{Uti}(p) = c) \wedge \mathsf{PreC}(p, d)) \rightarrow \mathsf{C}(p, c + d)]$

Precommitment by Elimination of Options Instead of relying on deliberative inertia, an agent may precommit 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. The idea is to take some drastic step that affects the cost of the actions or the utility of the intended conditions, and *then* to deliberate.

Cost Adjustment. An action is performed that leads the world along a scenario where the cost of the best plan to satisfy a commitment is higher than before. Interestingly, one cannot reduce the cost of the best plan through this technique, because (by Bellman's Principle) the best plan would automatically include any optimizations that might be available and known to the agent.

Example 14. In the refrigerator example of section 3.2, 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. ∎

D12. $\mathsf{A}[(\mathsf{PreC}(p, d) \wedge \mathsf{B}(\mathsf{Ecost}(\mathsf{Best}(p)) = c)) \rightarrow$
$(\bigvee a : \langle a \rangle \mathsf{true} \wedge \mathsf{B}(\langle a \rangle \mathsf{Ecost}(\mathsf{Best}(p)) = c + d) \wedge \langle a \rangle \mathsf{Delib}(x))]$

Utility Adjustment. Conversely, an agent may perform actions that would later make certain intentions more attractive, i.e., increase their utility to him then.

Example 15. 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 is formalized below. An agent with precommitment d for p performs an action after which his utility for p increases by d.

D13. $\mathsf{A}[(\mathsf{PreC}(p, d) \wedge \mathsf{B}(\mathsf{Uti}(p) = c)) \rightarrow$
$(\bigvee a : \langle a \rangle \mathsf{true} \wedge \mathsf{B}(\langle a \rangle \mathsf{Uti}(p) = c + d) \wedge \langle a \rangle \mathsf{Delib}(x))]$

5.3 Conative Policies and Deliberation

Generalizing our technical development, we can see that a natural extension would be to declaratively express *conative policies* describing not only how the agent acts and deliberates, but also how he controls his deliberations. These policies could accommodate not only the general rationality requirements studied

above, but also be made conditional upon properties of the domain, and the qualities of the agent, e.g., the qualities studied by [Kinny & Georgeff, 1991].

One can impose constraints on the conative policies of agents, e.g., to prevent them from adopting intentions that they believe are mutually inconsistent or inconsistent with their beliefs. The conative policies embodied in an agent would not change due to ordinary deliberations. Deliberations of a deeper nature would needed to create and modify them. These deliberations are a kind of "soul searching" that an agent may go through in deciding upon some value system. Intuitively, conative policies are commitments about commitments. Although, one can formally define arbitrarily nested commitments, we suspect that a small number of nestings would suffice in cases of practice, and in fact, greater nestings would probably prove counter-productive.

6 Conclusions and Future Work

We argued that commitments as well as the allied concepts of precommitments and conative policies are crucial for the design of intelligent, autonomous, but limited, agents. However, commitments must be formalized in a manner that is compatible with rationality, and which emphasizes the fact that real agents are limited. It is with this understanding that commitments are an effective component of a descriptive theory of rationality in limited agents. It is with the same understanding that they can prove useful in prescriptive theories as well.

There has been some good work in building experimental systems that capture the intentions, commitments, plans, and coordination strategies of agents, and in developing testbeds that can be declaratively customized to different situations, e.g., [Kinny & Georgeff, 1991; Sen & Durfee, 1994; Pollack *et al.*, 1994; Singh *et al.*, 1993]. However, a lot more good work is need to handle the kinds of examples and constraints we described above. Eventually, this work would need the expressiveness to handle the more general conative policies.

Future work includes formally expressing a wide range of conative policies that capture various rationality postulates as well as characterize interactions among agents. Real-time reasoning aspects were ignored above, but are important in several applications. Similarly, considerations on deliberating about and scheduling across multiple intentions are also important.

Overall, we believe the study of commitments in the architecture of limited, rational agents will prove highly fruitful.

References

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

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

[Brooks, 1991] Brooks, Rodney; 1991. Intelligence without reason. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).* Computers and Thought Award Lecture.

[Castelfranchi, 1995] Castelfranchi, Cristiano; 1995. Commitments: From individual intentions to groups and organizations. In *Proceedings of the International Conference on Multiagent Systems.*

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

[Emerson, 1990] Emerson, E. A.; 1990. Temporal and modal logic. In Leeuwen, J.van, editor, *Handbook of Theoretical Computer Science*, volume B. North-Holland Publishing Company, Amsterdam, The Netherlands.

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

[Kinny & Georgeff, 1991] Kinny, David N. and Georgeff, Michael P.; 1991. Commitment and effectiveness of situated agents. In *IJCAI.*

[McCarthy, 1979] McCarthy, John; 1979. Ascribing mental qualities to machines. In Ringle, Martin, editor, *Philosophical Perspectives in Artificial Intelligence.* Harvester Press. Page nos. from a revised version, issued as a report in 1987.

[Newell, 1982] Newell, Allen; 1982. The knowledge level. *Artificial Intelligence* 18(1):87–127.

[Norman & Long, 1996] Norman, Timothy J. and Long, Derek; 1996. Alarms: An implementation of motivated agency. In *Intelligent Agents II: Agent Theories, Architectures, and Languages.* 219–234.

[Pollack *et al.*, 1994] Pollack, Martha E.; Joslin, David; Nunes, Arthur; Ur, Sigalit; and Ephrati, Eithan; 1994. Experimental investigation of an agent commitment strategy. Technical Report 94-13, Department of Computer Science, University of Pittsburgh, Pittsburgh.

[Rao & Georgeff, 1991] Rao, Anand S. and Georgeff, Michael P.; 1991. Modeling rational agents within a BDI-architecture. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning.* 473–484.

[Sen & Durfee, 1994] Sen, Sandip and Durfee, Edmund H.; 1994. The role of commitment in cooperative negotiation. *International Journal of Intelligent and Cooperative Information Systems* 3(1):67–81.

[Simon, 1981] Simon, Herbert; 1981. *The Sciences of the Artificial.* MIT Press, Cambridge, MA.

[Singh & Asher, 1993] Singh, Munindar P. and Asher, Nicholas M.; 1993. A logic of intentions and beliefs. *Journal of Philosophical Logic* 22:513–544.

[Singh *et al.*, 1993] Singh, Munindar P.; Huhns, Michael N.; and Stephens, Larry M.; 1993. Declarative representations for multiagent systems. *IEEE Transactions on Knowledge and Data Engineering* 5(5):721–739.

[Singh, 1992] Singh, Munindar P.; 1992. A critical examination of the Cohen-Levesque theory of intentions. In *Proceedings of the 10th European Conference on Artificial Intelligence.*

[Singh, 1994] Singh, Munindar P.; 1994. *Multiagent Systems: A Theoretical Framework for Intentions, Know-How, and Communications.* Springer Verlag, Heidelberg, Germany.

[Singh, 1996] Singh, Munindar P.; 1996. A conceptual analysis of commitments in multiagent systems. Technical Report TR-96-09, Department of Computer Science, North Carolina State University, Raleigh, NC. Available at `http://www4.ncsu.edu/eos/info/dblab/www/mpsingh/papers/mas/commit.ps`.

A Formal Semantics

The semantics of $\mathcal{C}$ is given relative to intensional models. The formal model is as described informally in section 4. Let $M = \langle \mathbf{T}, <, [\![\,]\!], \mathbf{B}, \mathbf{C}, \mathbf{P}, \mathbf{A}, \Pi, \Omega \rangle$ be a model. Here $\mathbf{T}$ is a set of possible moments ordered by $<$; $[\![\,]\!]$ assigns intensions to atomic propositions, predicates, and actions. $\mathbf{B}$ assigns an alternativeness relation to each agent that capture his beliefs. $\mathbf{C}$, $\mathbf{P}$, and $\mathbf{A}$ assign a set of commitments, precommitments, and acting-for sentences to each agent at each moment. $\mathbf{S}_t$ is the set of scenarios originating at moment t. Π assigns a probability to each scenario at each moment, with the probabilities of the members of $\mathbf{S}_t$ adding up to 1. Ω assigns a utility (for each agent) to each condition at each moment.

The intension of an atomic proposition is the set of moments where it is true. The intension of a predicate is a function that takes yields a set of moments for each tuple of the predicate's arguments. The intension of an action is, for each agent x, the set of periods in which an instance of it is performed by x. Thus, $[t, t'] \in [\![a]\!]^x$ means that agent x performs action a from moment t to t'. We require that action instances be nonoverlapping. We lack the space to describe additional "coherence" (i.e., well-formedness) requirements.

The semantics of formulae of $\mathcal{C}$ is given relative to a model and a moment in it. $M \models_t p$ expresses "M satisfies p at t." $M \models_{S,t} p$ expresses "M satisfies p at moment t on scenario S," and is needed for scenario-formulae as defined in section 4.3. The satisfaction conditions for the temporal operators are adapted from those in [Emerson, 1990]. Formally, we have the following definitions:

S1. $M \models_t \psi$ iff $t \in [\![\psi]\!]$
S2. $M \models_t p \wedge q$ iff $M \models_t p \& M \models_t q$
S3. $M \models_t \neg p$ iff $M \not\models_t p$
S4. $M \models_t \mathsf{A}p$ iff $(\forall S : S \in \mathbf{S}_t \Rightarrow M \models_{S,t} p)$
S5. $M \models_t (\bigvee a : q)$ iff $(\exists b : b \in \mathcal{A}\ \&\ M \models_t q|_a^b)$, where $q|_a^b$ indicates the substitution of every occurrence of a by b in the expression q.
S6. $M \models_{S,t} p\mathsf{U}q$ iff $(\exists t' : M \models_{S,t'} q \& (\forall t'' : t \leq t'' \leq t' \Rightarrow M \models_{S,t''} p))$
S7. $M \models_{S,t} \mathsf{B}p$ iff $(\forall t' : t' \in \mathbf{B}(p) \Rightarrow M \models_{t'} p)$
S8. $M \models_{S,t} \mathsf{C}(p, c)$ iff $\mathsf{C}(p, c) \in \mathbf{C}_x(S, t))$
S9. $M \models_{S,t} \mathsf{PreC}(p, c)$ iff $\mathsf{PreC}(p, c) \in \mathbf{P}_x(S, t))$
S10. $M \models_{S,t} \mathsf{For}(a, p)$ iff $\mathsf{For}(a, p) \in \mathbf{C}_x(S, t) \& M \models_{S,t} \langle a \rangle \mathsf{true}$
S11. $M \models_{S,t} p$ iff $M \models_t p$, if semantic rules S6, S7, S8, S9, and S10 do not apply on p
S12. $M \models_t \mathsf{Ecost}(p) = e$ iff $\left(\sum_{S \in \mathsf{S}_t \& M \models_{S,t} \langle a \rangle \mathsf{true} \wedge \mathsf{Cost}(a) = c} \Pi_t(S) \times c \right) = e$
The expected cost of an action is the weighted sum of its costs along the scenarios on which it occurs.

This article was processed using the LaTeX macro package with LLNCS style