Unpublished manuscript, September 1991

Formalizing Rationality Postulates for Intentions

Munindar P. Singh*

ACT/Carnot Project and Dept of Computer Sciences
MCC, Austin, TX 78759, USA
University of Texas, Austin, TX 78712, USA

msingh@cs.utexas.edu

Abstract

Reasoning about intentions is an important issue in many subareas of AI. Therefore, it would be of great significance if one could express some rationality postulates that may be used to guide the design of intelligent agents who would deliberate effectively about their intentions and actions. However, even some simple kinds of such postulates lead to paradoxical results when attempted to be formalized directly. Here I propose an approach based on temporal possibility and action that avoids those paradoxes. This approach incorporates a novel model involving time and possibility in which a probabilistic analysis of *choice* can be formalized.

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

1 Introduction

Intentions and reasoning about intentions are a crucial part of many important subareas of AI—e.g., planning, plan recognition, situated action, and natural language understanding [6, 14]. 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 a certain intention [3, 12, 18]. However, before any theory of intentions can be applied to actual AI problems, we need to go beyond the semantics of intentions to consider how intelligent agents should reason with them. At least as a first approximation, it is useful to consider agents who are rational in some sense. Thus the problem we face reduces to that of stating some rationality postulates on the adoption, holding, and dropping of intentions by agents. These postulates are to be seen as essentially normative or prescriptive claims about how agents *ought* to reason (or ought to be designed to reason) with intentions.

It turns out that difficulties arise when these postulates are attempted to be formalized directly. Even some simple and uncontroversial intuitions about intentions lead to paradoxes, because they seem to be mutually incompatible. I submit that this happens because of the naive formalizations that one might attempt. I aim to show in this paper how rationality postulates on intentions can be perspicuously and non-paradoxically stated in a novel framework that also allows a rich language for making claims involving time, probability and choice.

The main motivation for presenting a probabilistic solution is that no purely qualitative solution seems to exist that is immune to the paradoxes discussed here, while a fairly simple probabilistic solution can be supplied. One could, of course, go all the way and bring in the concept of expected utilities. A solution based on expected utilities would work also, and might be preferable in cases where they are needed for other purposes anyway, e.g., in the approach used in [15] to formalize commitments, and in [17] to formalize precommmitments. However, the key intuitions that I wish to exploit for the arguments of this paper are all statable in a simple probabilistic framework. There is no need to introduce additional primitives such as utilities, since the resulting solution would have no advantages over the one proposed here. Therefore, on the methodological grounds of Occam's Razor, I use only probabilities below.

In §2, I informally discuss three problems that arise in formalizing some simple intuitions about intentions. In §3, I explain the ontological framework of this paper and point out its advantages over current frameworks. In §5, I present the formal language, and in §6, the formal model. Finally, in §7, I suggest formalizations of several rationality constraints on intentions. In particular, I return to the problems mentioned in §2, and solve each of them. The approach presented here is general and independent of the exact semantics given to intentions.

2 Three Problems

I now discuss the three paradoxes alluded to above, and their solutions, informally. This informal discussion is important since it motivates the technical approach of this paper.

2.1 Dudley Dolittle

One of the oldest problems in this context is due to McDermott [12, p. 102]: A movie hero, Dudley, sees his heroine, Nell, lying tied to a railroad track. He figures "Nell is going to be mashed" and decides to save her. Being a true hero, he is, of course, confident that he will succeed and therefore comes to believe that Nell will not be mashed, after all. From this he concludes that there is no more any need to do anything to save Nell. He might now realize again that Nell is in danger and go through the above reasoning repeatedly, or he might just pass time without further reasoning, but also without further action. In either case, the outcome would cost Dudley his reputation, and Nell her life.

Briefly, we want Dudley to realize that what happens in the world is not affected by merely intending—i.e., he must act so that the chances of Nell being mashed are in fact reduced. Dudley's first postulate that an intention leads to a belief in its success is well received in much of AI and Philosophy [12, p. 102], [14, p. 43], [3, 7, 18] and [2, p. 37] (who allows an "expectation," rather than a full belief). However, it can be safely weakened to allow a less confident agent than Dudley. Dudley's second postulate, namely, that he need not have an intention for something that is guaranteed is a special case of Les Lazy's postulate discussed below. We want to be able to retain some version of each of these postulates, and also to save Nell's life.

2.2 Les Lazy

It seems quite rational that an agent have an intention only if it is, in some way, useful for him to have it. That is, the intention must lead the agent to achieve something that not having the intention would not. For example, an agent should not intend something that (he believes) is going to happen anyway due to the environment, or due to his actions up to now. He should also not intend something that another agent is about to achieve. Les Lazy is one such agent. But under one formulation of this postulate, Les has an intention for a condition only if on all possible futures, his not having that intention would not guarantee that condition. Note that this is a cautious formulation since guarantees are considered, rather than contingent occurrences of the intended condition.

However, Les still ends up in trouble. He figures that if he intends q (where q entails p), then p would occur whenever q does. Thus even if he did not intend p, it would be assured on all futures where q is assured (which are the ones he is considering). This means that he cannot intend p. (This argument is discussed and formalized in §7.3.) Or, conversely, that he must drop his intention for q. The second option says that if Les intends p he cannot intend any proposition that entails it. This is exactly the opposite of *consequential closure* that plagues many logics of intention and belief, and is even more counterintuitive. But even the first option of not intending p could leave Les in the soup, e.g., if he passes up the only chance he would ever have to achieve p.

Briefly, the idea is that we want Les to act for each of his intentions. Thus even though q would lead to p, he might still do something to take care of p directly. This allows Les to take advantage of opportunities that arise as he acts. It also helps him ensure partial success in case his (possibly complex) plan for q fails. This is not the whole story, however—see §7.3 for postulates on actions.

2.3 Ken Klutz

Ken Klutz has learned from Dudley and Les that he must act for his intentions. But he is an incompetent agent who realizes that he cannot act properly. Indeed, he figures that if he intends p, the chances of p occurring are worse than if he never intended p. Then if he really intends to achieve p, he should *not* intend to achieve it! This paradox is adapted from Pears, and arises with any simple formalization of intentions [13].

The problem here is that we want Dudley and Les to act on their intentions, but do not want Ken to act on his, at least not immediately. Ken can have his intention, but actually act for it only when (and if) he believes that the chances of success would be improved. That is, he should wait until either p comes to hold for other reasons, or he gets an opportunity that even he cannot mess up, or p becomes so unlikely that it is wise even for him to act. Thus his intention should persist even though he should not aggressively act to achieve it.

3 Ontological Intuitions

For postulates such as the above to even be stated, we need a formal model that includes not just time and action, but also possibility and choice. Since I propose a probabilistic solution to the problems of §2, the model should also include some considerations of probability. Recently, a model for time, chance and action was proposed by Haddawy [8], who adapted it from the models that philosophers use for the semantics of conditionals—time is considered, but the *choices* made by an agent are not [19]. Thus this model, while interesting, has some shortcomings that obstruct its application to intentions, and indeed to the process of *choosing* actions.

The model I propose here is based on possible worlds. As diagramed in Figure 1, each possible *world* has a branching *history* of *times*. The actions an agent can do can differ at each time—this allows for learning and forgetting, and changes in the environment. 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." The construction so far has the same effect as Haddawy's (with respect to branching time), but we get his first two constraints for free [8, §4].

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" or "tomorrow." I shall not include these here, since they are irrelevant to the aims of this paper.

Even though a world may develop in several different ways, only one scenario can be actualized. I take probability as applying to scenarios in terms of *their chances of being actualized.* This probability is given relative to a world and time and some description of the 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.

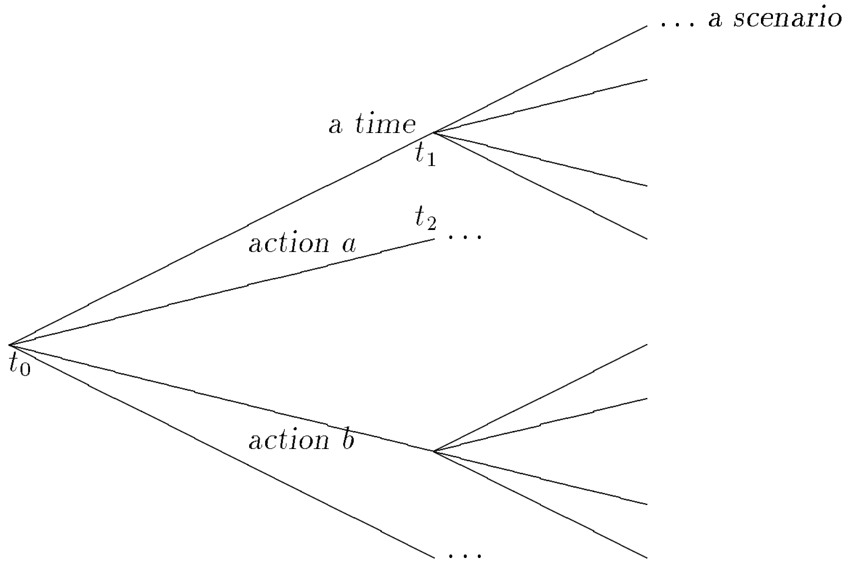

Figure 1: A World with a Branching History

Haddawy considers objective probability, but as "purely a function of the state of the world," and not dependent on the agent's internal state [8, §2]. This is stronger than van Frassen's requirement that only the *total history* (as against the current state) determines chance [19, p. 336]. But both have the problem that the agent's internal state is ignored altogether. This may be quite all right when only time and chance are involved, but is clearly too restrictive when actions are considered, since actions can influence what occurs later, and actions are chosen by agents depending on their internal state. In other words, in the proposed model, the *objective* conditional probability of the occurrence of a condition depends on the intentions and beliefs of the different agents.

Another useful feature of the formal language is quantification over basic actions. The idea is that of the several basic actions that an agent may possibly do at a given world and time, we often would like to restrictively talk of the subset of actions that have some interesting property. Indeed, we need something like this to formally express the idea of "choice"—an agent may be able to do several actions, but would, in fact, choose to do one (or a few that he can do simultaneously, but we can consider these formally as a single basic action), given the kind of agent he is.

While this is not crucial for the main claims of this paper, taking probability as applying to scenarios rather than propositions (as in [8, 11] [19, p. 334]) has some advantages. For one we can speak directly of the choice of actions by an agent as picking out a class of scenarios for actualization—an agent can *control* the world to some extent by constraining the scenario that would be actualized in this way. Of course, this class of scenarios is not usually a singleton, because an agent cannot control everything in the world—other agents and the environment also have a role to play in what ultimately transpires. With this approach, not only can we model choice perspicuously, we also get consistency since any scenario (a part of the model) must be internally temporally consistent. This idea also helps toward obtaining the values of probabilities: In most natural model-based reasoning, one can come up with the probabilities of scenarios directly, and must compute the probabilities for propositions indirectly. However, if one prefers, one can just as well take the probabilities as applying to propositions in the standard way.

4 Primitive Concepts

Clearly, we need the concepts of *intention* and *belief* as primitives—they are treated as predicates taking an agent and a proposition as arguments. The key feature of intentions that we need is that they lead to action. The sense of intentions that I am using is future-directed and allows 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 (given how he reasons) in order to satisfy an intention. The action an agent chooses to perform depends on the other intentions and beliefs he may have—the details of this are the concern of the planning community [6], and are not focused on here. Acting for an intention is a cognitive concept—it depends on the agent's internal state, rather than on 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 relate the agent's cognitive state with the world, we need the concept of *performing* an action in the world. I take it that an agent who acts for a condition intends it, and also immediately performs the action by which he acts for that condition.

Intentions and beliefs are given a simple semantics for ease of exposition, and to focus on the matters of interest—an *Intends*, *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 be analyzed using subjective probabilities as in [11, 19] for a more realistic analysis. *Performs* of an agent x and action a is true over a subscenario on which x does a.

Only objective probability is considered here and is treated as a function from propositions to the unit interval, $[0 \ldots 1]$. As explained in §3, in the model, probability is assigned to scenarios directly.

5 The Formal Language

The formal language of this paper, $\mathcal{L}$, is CTL* (a propositional branching time logic [4]), augmented with quantification over basic actions; a function, *Prob*; and predicates, *Believes*, *Intends*, *Acts-for* and *Performs* (as explained in §4). I also assume that all the arithmetic needed is available. CTL* is a logic that subsumes all propositional temporal logics [5], and so is ideal for our purposes in representing temporal knowledge—it allows all the qualitative statements we may wish to make. Standard CTL* (and, therefore, $\mathcal{L}$) are non-reified—this helps us make more natural claims about the past and future along scenarios. As Bacchus *et al.* have shown recently, non-reification is not a problem for representing temporal knowledge, so I shall not attempt to modify it [1].

$\mathcal{L}$ allows us to distinguish between present, past and future naturally, and to keep temporal considerations separate from probabilistic ones—this contrasts with Haddawy's logic in which past and future cannot be distinguished except by naming specific times, and the probability function incorporates an implicit future tense operator [8, §4]. By allowing quantification over scenarios, $\mathcal{L}$ also permits rationality postulates on intentions to be stated naturally.

Now I define the syntax of $\mathcal{L}$. Given a set of atomic formulae, Φ, $\mathcal{L}$ may be defined by the following rules.

1. Atomic formulae: $\phi \in \mathcal{L}$, for all $\phi \in \Phi$
2. Conjunction: $p, q \in \mathcal{L} \Rightarrow p \wedge q \in \mathcal{L}$
3. Negation: $p \in \mathcal{L} \Rightarrow \neg p \in \mathcal{L}$
4. Until: $p, q \in \mathcal{L} \Rightarrow p\mathsf{U}q \in \mathcal{L}$
5. Past: $p \in \mathcal{L} \Rightarrow \mathsf{P}p \in \mathcal{L}$
6. Path-quantifier: $p \in \mathcal{L} \Rightarrow \mathsf{A}p \in \mathcal{L}$

These formulae are given the usual semantics, which is formally defined in §6. A is a "path-quantifier." $\mathsf{A}p$ denotes "in *all* scenarios at the present time, p holds." Another path-quantifier is E. $\mathsf{E}p$ denotes "in *some* scenario at the present time, p holds"—i.e., $\mathsf{E}p \equiv \neg\mathsf{A}\neg p$. U stands for "until." $p\mathsf{U}q$ denotes that q holds sometimes on the given scenario and p holds until then. $\mathsf{F}p$ 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 "$\neg\mathsf{F}\neg p$." $\mathsf{P}p$ denotes "p held at some point in the past." Implication ($p \rightarrow q$) and disjunctions of formulae ($p \vee q$) are defined as the usual abbreviations.

Examples of wff's of $\mathcal{L}$ are (1) $\mathsf{AG}p$ "p holds at all possible times in the future," (2) $\mathsf{AF}p$ "p is inevitable," (3) $\mathsf{EF}p$ "p will occur in some future," (4) $\mathsf{AG}(\mathit{Acts\text{-}for}(x,a,p) \rightarrow \mathsf{A}(\mathit{Performs}(x,a) \rightarrow \mathsf{F}p))$ "if x does a for p, then all scenarios where he does a, p occurs eventually"—i.e., x's plan is perfect and a is all he needs to do for p, (5) $\mathsf{AG}(\mathit{Intends}(x,p) \rightarrow (\exists a : \mathit{Acts\text{-}for}(x,a,p) \wedge \mathit{Performs}(x,a)))$ "if x intends p then there is an action he can do for it," and (6) $\mathsf{AG}(\mathit{Intends}(x,p) \wedge \mathit{Acts\text{-}for}(x,a,p)) \rightarrow \mathit{Believes}(x, \mathit{Prob}(\mathsf{F}p|\mathit{Performs}(x,a)) > \mathit{Prob}(\mathsf{F}p))$ "if x intends p and decides to do a for p, then he believes that doing a would improve the chances of p occurring eventually." Note that not all of these formulae would be true in all cases.

6 The Formal Model

The semantics of $\mathcal{L}$ is given relative to intensional models: it is standard for CTL*, but is repeated here for completeness. The formal model is as described informally in §3. 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. $\mathbf{B}$ 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.

Formally, scenarios are defined using the concept of a *skeletal scenario*, which is an eternal linear sequence of times in the future of the given time; i.e., SS at w, t is any sequence: $\langle t = t_0, t_1, \ldots \rangle$, where $(\forall i : i \geq 0 \rightarrow t_i < t_{i+1})$ (linearity) and $(\forall i, t' : t' > t_i \rightarrow (\exists j : t' \not> t_j))$ (eternity). Now, a scenario, S, for w, t is the "linear closure" of some skeletal scenario at w, t. Formally, S, relative to some SS, is the minimal set that satisfies the following conditions:

- *Eternity*: $SS \subseteq S$
- *Linear Closure*: $(\forall t'' : t'' \in S \rightarrow (\forall t' : t_0 < t' < t'' \rightarrow t' \in S))$

$\mathbf{S}_{w,t}$ 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.

Basic actions may have different durations relative to one another in different scenarios, even those that begin at the same world and time. Here the basic actions always succeed; the model can be extended to allow failure. 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 (as models of actions), we need to constrain them so that (1) at any time, each action can be done in at most one way on any given scenario; (2) subscenarios are uniquely identified by the times over which they stretch, rather than the scenario used to refer to them; (3) there is a always future time available in the model; and (4) something must be *done* by each agent along each scenario in the model, even if it is some kind of a dummy action.

Restrictions on $[\![\,]\!]$ can also be used to express the limitations of agents, and the ways in which their actions may interfere with those of others; e.g., x cannot pick up three glasses at once and at most one person can enter an elevator at a time. The habits of agents can be similarly modeled; e.g., x always brakes before turning.

The semantics of sentences (i.e., formulae) in the formal language 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 path-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 [4]. 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. $M \models_{w,t} \mathsf{A}p$ iff $(\forall S : S \in \mathbf{S}_{w,t} \rightarrow M \models_{S,t} p)$

5. $M \models_{S,t} p\mathsf{U}q$ 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.

6. $M \models_{w,t} \mathsf{P}p$ iff $(\exists t' : t' < t \wedge M \models_{w,t'} p)$
7. $M \models_{S,t} p$ iff $M \models_{w,t} p$, if p is not of the form $q\mathsf{U}r$, and w is the unique world such that $S \in \mathbf{S}_{w,t}$
8. $M \models_{S,t} \mathit{Believes}(x,p)$ iff $(\exists t' : \mathit{Believes}(x,p) \in \mathbf{C}_x(\langle S,t,t'\rangle))$
9. $M \models_{S,t} \mathit{Intends}(x,p)$ iff $(\exists t' : \mathit{Intends}(x,p) \in \mathbf{C}_x(\langle S,t,t'\rangle))$
10. $M \models_{S,t} \mathit{Acts\text{-}for}(x,a,p)$ iff $(\exists t' : \mathit{Acts\text{-}for}(x,a,p) \in \mathbf{C}_x(\langle S,t,t'\rangle))$
11. $M \models_{S,t} \mathit{Performs}(x,a)$ iff $(\exists t' : \langle S,t,t'\rangle \in [\![a]\!]^x)$
12. $M \models_{w,t} \mathit{Prob}(p) = r$ iff $[\sum_{S \in \mathbf{S}_{w,t} \wedge M \models_{S,t} p} \pi_{w,t}(S)] = r$

    The probability of p holding is r iff the probabilities of all the scenarios that satisfy it add up to r.

7 Postulates for Intentions

I have presented a general framework for making statements about intentions, actions and related concepts, which I have claimed has all the features we need to be able to express postulates on intentions non-paradoxically. I now turn to these postulates, and discuss and formalize them one by one.

7.1 Relationship with Beliefs and Other Intentions

1. Intended propositions must be believed possible.

   $\mathsf{A}(\mathit{Intends}(x,p) \rightarrow \mathit{Believes}(x, \mathsf{EF}p))$

2. Intended propositions must not be believed inevitable.

   $\mathsf{A}(\mathit{Intends}(x,p) \rightarrow \neg\mathit{Believes}(x, \mathsf{AF}p))$

Though these conditions do not involve actions, they are sufficient to block the kinds of inferences that are needed for Dudley Dolittle's paradox to arise. This is because Dudley's intention just leads to a belief that he may succeed, but he is able to give up his intention only if he believes that success is inevitable, which is a much stronger condition.

However, these conditions are simply too weak to be of much help in practice. This is because too many of the naturally arising conditions that our agents would intend would be believed by them to be possible anyway—i.e., postulate 1 is usually redundant. More useful formalizations of the intuitions behind it are presented in postulates 3, 5, 6 and 7.

And while postulate 2 is useful in telling us when intentions ought to be dropped by a rational agent, it applies only if the agent believes p to be inevitable, which is quite a strong requirement—ordinarily, none of the conditions an agent would seriously intend might ever be believed inevitable by him (until they finally occur). Postulate 2, however, represents a useful intuition that is captured by the final version of the postulates proposed in this paper.

7.2 Acting for an Intention

It might seem that incorporating actions explicitly would help us solve Les Lazy's problem. When we bring in actions, a more fine-grained postulate comes immediately to mind: An intention should entail that the agent holding it believe it possible that he will act for it. This is surely too weak to be controversial, for what would be the intentions of an agent who did not even consider it possible that he will act for them. This postulate is supposed to further capture the intuition that having an intention should lead to something different than what not having it would, although it allows fortuitous circumstances under which the agent might not have to act. It is easy to express formally:

3. Intentions must be believed possible to act for.

   $\mathsf{A}(\mathit{Intends}(x, p) \rightarrow \mathit{Believes}(x, \mathsf{EF}(\exists a : \mathit{Acts\text{-}for}(x, a, p))))$

This inference is quite weak, and helps us estimate only a little of an agent's doxastic state concerning his future actions. All that this inference gives us is the fact that an agent who has an intention thinks that he can act for it, at least on some future. One way of strengthening this and subsequent postulates is by using the notion of *know-how* as developed in [16]—an agent should, under some circumstances, know how to achieve his intention. However, there is no need for this complication here.

Instead, let us consider the postulate that an agent who intends something has an action that he performs for that intention. This would hold if, e.g., the agent had a currently adopted plan corresponding to his intention. This postulate is apparent in the suggestion of Georgeff that intentions be treated as "adopted goals" [6], and is sufficiently strong to capture the relationship between intentions and actions, e.g., to make Dudley do something to save Nell's life.

4. Intentions are "adopted goals;" i.e., they must be acted for immediately.

   $\mathsf{A}(\mathit{Intends}(x, p) \rightarrow (\exists a : \mathit{Acts\text{-}for}(x, a, p)))$

7.3 The Point of Intending

I now consider the intuitive justifications for Les Lazy's postulate and look at ways to express it formally. Les's postulate is stronger, and seemingly more reasonable, than postulate 2. An agent can have an intention only if he believes that if he did not have this intention, the intended condition would not come to hold. That this seems rather strong is a consequence of the fact that I am assuming a two-valued logic here; when degrees of belief or subjective probabilities are brought in, a more reasonable condition may be expressed. However, that issue is orthogonal to the main concern of this paper.

Note that the current formal language suffices to express the kind of counterfactual conditionals needed to express this postulate, since they can be treated as conditionals involving quantification over scenarios. The antecedent is the not having of the given intention (by the given agent), and we quantify over all scenarios on which this is the case. Note that it does not even make sense to state that if the agent does not have the intention the given condition would not be possible at all (i.e., $\neg\mathsf{EF}p$), since this is evaluated at world-time pairs in the model. If it is true at a world-time pair, it is true there even if the agent does not have the intention. Instead, we must treat the counterfactual as a path formula, i.e., as one that would be evaluated on scenarios rather than at world-time pairs. Some interesting options are considered below.

5. An agent intends something only if he believes that on all scenarios (at that time) where he does not have the intention, the intended condition does not ever occur.

   $\mathsf{A}(\mathit{Intends}(x,p) \rightarrow \mathit{Believes}(x, \mathsf{A}(\neg \mathit{Intends}(x,p) \rightarrow \neg \mathsf{F}p)))$

6. An agent intends something only if he believes that on all scenarios (at that time) where he does not have the intention, the intended condition is not guaranteed, i.e., is not inevitable.

   $\mathsf{A}(\mathit{Intends}(x,p) \rightarrow \mathit{Believes}(x, \mathsf{A}(\neg \mathit{Intends}(x,p) \rightarrow \neg \mathsf{FAF}p)))$

Postulate 5 requires that the agent believe that on all scenarios where he gives up his intention for p, the intended condition simply would not occur, i.e., not even accidentally. This sort of a requirement can make sense only for conditions that are fully under the control of the given agent, so that if he does not actively intend that they hold, they do not hold—i.e., they cannot come to hold because of events in the environment or because of the actions of other agents. Interestingly, there are natural conditions such as these. For example, an agent's speech acts would ordinarily be taken as being entirely within his control. However, most kinds of conditions in the world are not like this. For example, if the agent does not open a particular door, someone else might, or it might open because of the breeze. But even for such conditions, while this postulate is not entirely true, it might be defeasibly acceptable. For example, if the door is locked and only the agent has the key, then it should not become open unless the agent intends it to—at least, it might be reasonable for the agent to assume this defeasibly.

In general, however, postulate 5 is too strong. Postulate 6 is significantly weaker and, therefore, more easily acceptable. It states that while the intended condition might occur even on scenarios where the agent does not have the given intention, it cannot be guaranteed to occur on those scenarios. On such scenarios the intended condition might still hold due to extraneous events, but the agent should not believe that it is guaranteed. For example, a door might open even if the agent does not intend to open it, but that could be because of the environment, or other agents, or some incidental effects of the agent's own actions.

In either case, however, the counterfactual can easily lead to trouble. This is because it validates the following condition, which is nothing but Les Lazy's problem, as described in §2.2.

- $\mathsf{A}(\mathit{Intends}(x,p) \wedge \mathit{Believes}(x, \mathsf{AG}(q \rightarrow p)) \rightarrow \neg \mathit{Intends}(x,q))$

  or put slightly differently:

- $\mathsf{A}(\mathit{Intends}(x,q) \wedge \mathit{Believes}(x, \mathsf{AG}(q \rightarrow p)) \rightarrow \neg \mathit{Intends}(x,p))$

What we have here might be termed *consequential anti-closure:* a most unusual problem for a logic! Since we clearly do not want to limit application of our theory to agents who are never able to make the inference used in the above, we must give up or reformulate the postulate that led to this problem.

Another problem occurs if the agent does not have the intention that is being considered right now, but is about to adopt it (or one for a condition that entails it). This should be taken care of naturally, provided that there are scenarios still possible on which the agent does not adopt this intention. This is because the above postulates involve an A on scenarios, so if there are some scenarios where the stated condition ($\mathsf{F}p$ or $\mathsf{AF}p$) does not hold, the postulate might still be satisfied.

One can try an explanation or justification for consequential anti-closure along the following lines. Let q entail p. If x intends p and also intends q (for different reasons), we can avoid the bad effects of this inference by requiring it to mean functionally that x has only one intention *active*. If he succeeds with both, so much the better; if he fails with the intention that was active but finds that the other is still possible (and there are still reasons for him to have it), he can continue to try it. However, this proposal seems quite unnatural and contrived. There is no real reason for making our agents act sequentially: this is an improper restriction especially where the conditions intended are deadline bound or where the agent is internally a multiagent system, different members of which may simultaneously act on its different intentions. In other words, we should allow our agents to simultaneously act for more than one intention.

7.4 Commitments

Instead of trying to work around consequential anti-closure, I will try to formalize the above requirement differently. One idea is to fall back upon the notion of *commitment* as defined by Harman [9, p. 94] and Bratman [2, ch. 2]. An agent has an intention only as long as he is committed to it. As before, let q entail p. Intuitively, if the agent acts for q on the basis of his reasons for q, he may achieve not only achieve q, but also achieve p as a side-effect (in fact, he might achieve p even if cannot achieve q). But even if he fails in achieving q and does not achieve p because of it and gives up his intention for q altogether, he would still have his original reasons intact for intending p. Thus he could still intend p, and possibly achieve it. Note that the commitments for both p and q are held concurrently—no sequencing is implied.

A natural way of treating commitments formally is by using probabilities. Each commitment somehow improves the chances of success of the corresponding intention. Therefore, the probability of p occurring in the future increases when q is intended, but increases some more when p itself is also intended. Thus, in probabilistic terms, the reasonable intuition behind postulate 5 may be expressed as in postulate 7 below. The difference between postulate 5 and postulate 6 may be safely neglected for our purposes here. That is, postulate 7 replaces postulate 5 and 6.

7. An agent believes that his intention raises the probability of success.

$$\mathsf{A}(\mathit{Intends}(x,p)\rightarrow\ \mathit{Believes}(x, \mathit{Prob}(p|\mathit{Intends}(x,p)) - \mathit{Prob}(p|\neg\mathit{Intends}(x,p)) > \epsilon))$$

Here $\epsilon > 0$ is just a parameter telling us how much likelier is likelier enough. This says that x intends p only if x believes that p is likelier if x intends p than otherwise. Thus even if x intends q and believes that q entails p in the future, he can intend p, if intending p would make it likelier for p to be achieved than if just q were intended. Note that here q is not intended because it would make p likelier, but only for itself, if at all. This postulate finally takes care of Les's difficulties.

But in its present form, it does not take care of Ken Klutz, whose problem is that the probability of his achieving p is worse if he intends p than otherwise. The above intuition about commitments can be used for this also, if we realize that while an agent may be committed for acting for his intentions, he does not have to be committed to doing so now, only *eventually*, i.e., when the moment is right.

8. An agent with an intention will act for it eventually, if it does not come to hold on its own. This postulate replaces postulate 4, which is one of the sources of Ken's mistake.

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

9. An agent with an intention acts for it only when he believes that doing so would raise the probability of succeeding (significantly). This postulate replaces postulate 7, which is the other source of Ken's mistake. It is not just having an intention, but acting for it that should raise the probability of success; i.e., the agent should act only if he needs to, and as we can see from postulate 8, an agent might not always need to act for his intentions.

   $\mathsf{A}(\forall a : \mathit{Intends}(x,p) \wedge \mathit{Acts\text{-}for}(x,a,p) \rightarrow$
   $\mathit{Believes}(x, \mathit{Prob}(p|\mathit{Acts\text{-}for}(x,a,p)) - \mathit{Prob}(p|\neg \mathit{Acts\text{-}for}(x,a,p)) > \epsilon))$

It is clear that we need at least the postulates 8 and 9. We can include either one of postulates 2 or 3, since either combined with postulates 8 and 9 would entail the other. For concreteness, I adopt postulate 3. This entails postulate 2, since it requires that the agent with an intention believe he will act for it. Since $\epsilon > 0$, he will act for it only if it strictly increases the probability of the occurrence of the intended condition, which means that the intended condition cannot have been believed inevitable. Again, since $\epsilon > 0$, postulate 9 entails that, if the agent intends p (and acts for it), he believes that the probability of its occurrence is non-zero. In other words, if he intends p and acts for it, he believes that it will occur on some future. Thus postulate 9 entails postulate 1 under the assumption that intentions are (believed to be) eventually acted for on at least some possible future. But this assumption is precisely what postulate 3 captures. Therefore, the latter does not have to be separately stated. Thus we have non-paradoxical formulations of the intuitions behind the postulates on intentions, which when naively formalized lead to problems.

8 Conclusions

I have presented a novel framework for representing intentions, actions and beliefs that involves probability and choice, and is both powerful and elegant. It allows us to express some intuitive, but troublesome, rationality postulates on intentions most naturally and to dissolve three putative paradoxes of intentions. The intuitions captured by these postulates can be easily carried over to utility-theoretic versions of them, e.g., by replacing changes in probability by changes in utility. Such a move would be necessary when one tries to give a more concrete account of how limited rational agents adopt and drop their intentions [15, 17]. Postulates for planning may also be stated in the same general scheme—not only would such postulates be a step toward relating rationality with planning, they would also provide a sound basis for agents to infer the plans of others.

References

[1] Fahiem Bacchus, Josh Tenenberg, and Johannes A. Koomen. A non-reified temporal logic. In *First Conference on Knowledge Representation and Reasoning*, pages 2–10, 1989.

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

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

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

[5] E. Allen Emerson and Joseph Y. Halpern. "Sometimes" and "Not Never" Revisited: On Branching versus Linear Time Temporal Logic. *Journal of the ACM*, 33(1):151–178, 1986.

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

[7] H. Paul Grice. Intentions and uncertainty. In *Proceedings of the British Academy*, pages 263–279, 1971.

[8] Peter Haddawy. Time, chance, and action. In *Sixth Conference on Uncertainty in AI*, 1990.

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

[10] William L. Harper, Robert Stalnaker, and Glenn Pearce, editors. *IFS: Conditionals, Belief, Decision, Chance, and Time.* D. Reidel, Dordrecht, Netherlands, 1981.

[11] David Lewis. A subjectivist's guide to objective chance. In *[10]*, pages 267–297. 1981.

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

[13] D. F. Pears. Intention and belief. In Bruce Vermazen and Merrill Hintikka, editors, *Essays on Davidson: Actions and Events.* Oxford University Press, Oxford, UK, 1985.

[14] Martha E. Pollack. *Inferring Domain Plans in Question Answering.* PhD thesis, University of Pennsylvania, 1986.

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

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

[17] Munindar P. Singh. On the commitments and precommitments of limited agents. In *IJCAI Workshop on the Theoretical and Practical Design of Rational Agents*, August 1991.

[18] Munindar P. Singh and Nicholas M. Asher. A logic of intentions and beliefs. *Journal of Philosophical Logic*, 1992. In press.

[19] Bas C. van Frassen. A temporal framework for conditionals and chance. In *[10]*, pages 323–340. 1981.