Semantical Considerations on Intention Dynamics for BDI Agents

Munindar P. Singh[*]
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-8206, USA

singh@ncsu.edu
+1.919.515.5677

October 24, 1997

Abstract

The BDI paradigm is a powerful means for constructing intelligent agents in terms of their *b*eliefs, *d*esires, and *i*ntentions. For this paradigm to bear its full potential, it must incorporate considerations from rationality. This paper develops a set of postulates for intelligent agents who deliberate about their intentions and actions. However, even simple postulates can lead to paradoxical results when formalized naively. We propose an approach based on temporal possibility and action that avoids those problems. This approach incorporates a formal model based on branching time in which a probabilistic analysis of *choice* can be captured. In this manner, the intuitions of the BDI paradigm can be reconciled with those of rational agency.

Short title: Semantical Considerations on Intention Dynamics

[*] Munindar Singh is supported by the NCSU College of Engineering, the National Science Foundation under grants IRI-9529179 and IRI-9624425 (Career Award), and IBM corporation.

Semantical Considerations on Intention Dynamics for BDI Agents

(Author's name removed as asked for in the instructions)

Abstract

The BDI paradigm is a powerful means for constructing intelligent agents in terms of their *b*eliefs, *d*esires, and *i*ntentions. For this paradigm to bear its full potential, it must incorporate considerations from rationality. This paper develops a set of postulates for intelligent agents who deliberate about their intentions and actions. However, even simple postulates can lead to paradoxical results when formalized naively. We propose an approach based on temporal possibility and action that avoids those problems. This approach incorporates a formal model based on branching time in which a probabilistic analysis of *choice* can be captured. In this manner, the intuitions of the BDI paradigm can be reconciled with those of rational agency.

1 Introduction

The BDI model is a powerful means to understand, study, and design intelligent agents [Georgeff and Rao, 1995]. The BDI model presents an abstract architecture of intelligent agents in terms of their *b*eliefs, *d*esires, and *i*ntentions. BDI concepts apply in a number of areas, including planning and natural language understanding [Pollack, 1992; Grosz and Sidner, 1990; Breiter and Sadek, 1996]. The key research challenges in the BDI approach are

(a) *semantics:* how to relate the physical states of an agent with its beliefs, desires, and intentions,

(b) *actions:* how to relate these concepts to the agent's actions, and

(c) *evolution:* how to constrain the development of the agent's state.

These challenges are crucial to any attempt to build effective BDI agents. Challenge (a) has drawn the most research attention; this paper focuses on important aspects of challenges (b) and (c), which we jointly label *dynamics.* Understanding the dynamics is essential for designing BDI agents that operate in a changing world.

Of the BDI concepts, belief has been the most intensively studied. Desires are typically treated as given. There is broad agreement that beliefs and intentions are subject to consistency and rationality requirements, whereas desires are not. Intentions mediate between an agent's desires and beliefs and relate most immediately to actions [Bratman, 1987]. They are also more complex than beliefs. For these reasons, we focus primarily on intentions here. Most work on the formalization of intentions is logical or qualitative, not quantitative [Cohen and Levesque, 1990;

Rao and Georgeff, 1991; Singh, 1994]. It addresses the semantics challenge, and has a descriptive, rather than a normative focus. Some existing work considers the normative aspect of intentions, but primarily in a qualitative framework—section 5 contrasts this work with our approach. By contrast, this paper develops a normative, quantitative view of BDI dynamics, albeit in a logical framework.

It is often useful, and indeed customary, to assume that agents are *rational* [Doyle, 1992]. They may not be perfect reasoners, but use their bounded resources as best they can. Rationality is essential to understanding how agents *should* act. Incorporating rationality in intention dynamics is thus crucial to the success of the program of research into BDI architectures.

Our approach takes the form of *postulates* on when intentions must be adopted, held, and dropped. These postulates are normative claims about how BDI agents *ought* to reason with intentions. Interestingly, difficulties arise in formalizing the requisite postulates qualitatively. Even some simple and uncontroversial intuitions about intentions appear mutually incompatible, if not paradoxical. However, when the traditional qualitative models are augmented with probabilities, the required rationality postulates can be consistently stated.

Doing so turns out to be nontrivial. On beginning this research, we encountered two major conceptual difficulties, both from unexpected quarters.

- The formal semantics of intentions as given by the existing theories is inherently static—intentions are evaluated at states, whereas the present task demands dynamic definitions—intentions should be evaluated along paths or histories. The problem is not so much with the details of these proposals as in their static formulation. Consequently, we formulate intentions and beliefs as "path formulae" [Emerson, 1990]. This can be adapted to most proposals.

- The probabilistic theories of belief and action are restrictive. One can either deal with

  - *subjective probabilities*, which depend exclusively on the internal state of an agent, or
  - *objective chance*, which depends exclusively on the state of the world.

  For example, [Haddawy, 1996, p. 3] considers objective chance, dependent on the state of the world, and not on the agent's state. This is stronger than van Fraassen's requirement that the *total history* (as opposed to just the current state) determines chance [van Fraassen, 1981, p. 336]. However, the objective chance depends on *both* the state of the world and the internal state of the agent. The agent's intentions determine his actions, and his actions influence what objectively transpires. Knowing that you intend to mow the lawn, one might conclude that the probability of the lawn being mowed is higher than otherwise, but this improvement is not in the mind of any agent—it is an objective improvement!

When these conceptual difficulties became clear, we were able to develop a simple framework with a formal language for making claims involving time, probability, and choice. The results of this paper could also be formulated using decision-theoretic notions. However, since simple probabilities suffice for our immediate needs, this paper does not invoke those more complex notions.

Section 2 informally discusses some possible postulates that capture some simple intuitions about intentions, and introduces three representative problems that arise in formalizing them. Section 3 describes our technical framework. Section 4 presents several postulates and discusses their ramifications. It offers a small set of refined postulates that solve the problems of section 2.

2 Problems in Intention Dynamics

For motivation, we discuss three apparently paradoxical situations involving how BDI agents may act on, and revise, their intentions. These situations arise when some intuitive constraints are applied naively. These situations are by no means claimed to be all of the problematic situations that can arise when reasoning about intentions. They are chosen as a simple benchmark for theories of intention dynamics, because they urge postulates that are apparently contradictory—postulates that we argue cannot be expressed in a purely qualitative framework. These problems are thus a minimal set justifying the marriage of logical and probabilistic approaches to reasoning about intentions.

2.1 Dudley Dolittle

A movie hero, Dudley, sees his heroine, Nell, lying tied to a railroad track [McDermott, 1982, p. 102]. He figures "Nell is going to be mashed" and adopts an intention to save her. Being a hero, he is 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 a need to do anything to save Nell. Thus he drops his intention to save Nell. He might now realize again that Nell is in danger and go through the above reasoning repeatedly, or forget about it altogether.

Briefly, Dudley should realize that what happens in the world is not affected merely by 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, e.g., [Bratman, 1987, p. 37]. 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. Ideally, one would like to retain some version of each of these postulates, but also to prevent the premature dropping of intentions.

2.2 Les Lazy

Arguably, an agent should have an intention only if it is 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. 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.

However, Les still ends up in trouble. Let q be the proposition that Les obtains a pen, and p be the proposition that Les obtains a pen or a pencil. Les knows that q entails p. He figures that if he intends q, then even if he did not intend p, it would be assured on all futures where q is assured. This means that he cannot intend p. Conversely, if Les intends p, he cannot intend any proposition that entails p.

Briefly, what one would like is that Les should 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 would allow Les to take advantage of opportunities that arise as he acts. It would also help him obtain partial success in case his plan for q fails.

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 generalizes the problem of *akrasia*, which occurs when an agent cannot act for his intentions [Pears, 1985]. An akrastic agent cannot do any good, but Ken is worse: he causes harm!

The theory must ensure that whereas Dudley and Les should act on their intentions, Ken should not act on his, at least not immediately. Ken can have his intention, but defer acting for it until (if ever) 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. In fact, we will later argue that there must be some circumstances in which the agent's actions for an intention improve its chances. Thus his intention should persist even though he should not aggressively act to achieve it.

3 Technical Framework

For postulates such as the above to even be stated requires a formal model that includes not just time and action, but also possibility and choice [Chellas, 1992]. A number of good approaches exist in the literature; for concreteness, we base our model on the presentation in [Emerson, 1990].

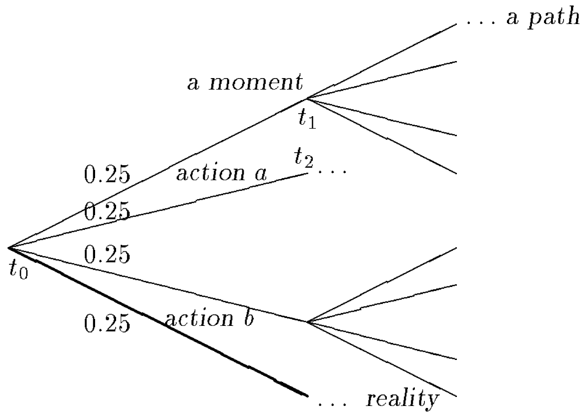

Figure 1: The Formal Model

Our formal model consists of moments, each describing a possible state of affairs and a possible cognitive state of the agent (a single agent named x). As diagramed in Figure 1, each possible *moment* can evolve in a number of possible ways depending on the actions of our chosen agent, the actions of other agents, and events in the environment. Thus in Figure 1, our agent can perform either action a or action b starting in moment t_0. If he does action a, the resulting moment could be either t_1 or t_2, depending on what else happens concurrently. For expository ease, we assume linear past: *periods* are defined by their start and end moments. Thus, $[t_0, t_1]$ denotes the period from t_0 through t_1, which includes all moments (if any) that lie in between. We assume that each agent does exactly one action over any period.

A *path* is any single branch of time that begins at the given moment. The set of paths emerging from moment t is given by $\mathbf{P}_t$. All of these paths have t as their initial moment. Given a path P and a moment t' on it, $P \downarrow t'$ denotes the subpath or suffix of P that begins at (and includes) t'. $P^0 = t$ denotes the initial moment in P_t.

Although several paths may emerge from each moment, one path out of each moment will be realized. For example, in Figure 1, the bottom path out of t_0 is the real path. Our postulates can be thought of as constraints on which path becomes the real one.

Primitives. We take probability as applying to paths in terms of *their chances of being realized.* It can be conditionalized on descriptions of an agent's cognitive state, i.e., his beliefs, desires, and intentions. Consequently, even for the same moment, different cognitive states can be considered. This enables us to express probability with and without a particular intention—crucial to this paper. Intuitively, the path that turns out to be the real path only becomes known after the fact.

For simplicity, beliefs and intentions are treated as modal operators β and ι. We require that agents have true beliefs about their beliefs and intentions. Performing an action is an operator δ whose argument is an action. We introduce a primitive, α, which means *acting for* (an intention), which applies to an action and a condition. This helps relate intentions to actions—we require that the action is in fact performed and the condition is intended. An agent may act for an intention even if it would be impossible or unlikely to succeed. The same action could be performed for different intentions; of course, several distinct and temporally isolated actions may have to be performed for a single intention.

Syntax. The formal language of this paper, $\mathcal{R}$, is CTL* (a propositional branching time logic [Emerson, 1990]) with some augmentations. Let Φ be a set of atomic formulae; and $\mathcal{A}$ is a set of action symbols. $\mathcal{R}$ may be defined by the following rules, which for ease of exposition simplify some of the structural aspects of the syntax of CTL* as given in [Emerson, 1990].

R1. Atomic formulae: $\phi \in \mathcal{R}$, for all $\phi \in \Phi$

R2. Conjunction: $p, q \in \mathcal{R} \Rightarrow p \wedge q \in \mathcal{R}$

R3. Negation: $p \in \mathcal{R} \Rightarrow \neg p \in \mathcal{R}$

R4. Until: $p, q \in \mathcal{R} \Rightarrow p\mathsf{U}q \in \mathcal{R}$

R5. Path-quantifier: $p \in \mathcal{R} \Rightarrow \mathsf{A}p \in \mathcal{R}$

R6. Action-quantifier: $p \in \mathcal{R}$, $a \in \mathcal{A} \Rightarrow (\bigvee a : p) \in \mathcal{R}$

R7. Belief: $p \in \mathcal{R} \Rightarrow \beta p \in \mathcal{R}$

R8. Intention: $p \in \mathcal{R} \Rightarrow \iota p \in \mathcal{R}$

R9. Acts-for: $p \in \mathcal{R}$, $a \in \mathcal{A} \Rightarrow \alpha(a, p) \in \mathcal{R}$

R10. Performs: $a \in \mathcal{A} \Rightarrow \delta a \in \mathcal{R}$

R11. Probability: $p \in \mathcal{R}$, $r \in [0..1] \Rightarrow (Pr(p) = r) \in \mathcal{R}$

Intuitively, A is a "path-quantifier." $\mathsf{A}p$ denotes "in *all* paths at the present moment, p holds." $\mathsf{E}p$ abbreviates $\neg\mathsf{A}\neg p$. Thus, $\mathsf{E}p$ denotes "in *some* path at the present moment, p holds." U stands for "until." $p\mathsf{U}q$ denotes that q holds sometimes on the given path and p holds until then. $\mathsf{F}p$ denotes "p holds sometimes in the future on this path" and abbreviates "$\mathsf{true}\mathsf{U}p$." $\mathsf{G}p$ denotes "p always holds in the future on this path" and abbreviates "$\neg\mathsf{F}\neg p$." $(\bigvee a : p)$ means that there is an action, which when substituted into p evaluates to true. $\bigwedge$ is the dual of $\bigvee$. Implication $(p \rightarrow q)$ and disjunctions of formulae $(p \vee q)$ are defined as the usual abbreviations.

Example 1 Examples of well-formed formulae of $\mathcal{R}$ are

1. $\mathsf{AG}p$ "p holds at all possible moments in the future"

2. $\mathsf{AF}p$ "p is inevitable"

3. $\mathsf{EF}p$ "p will occur on some path in the future"

4. $\mathsf{AG}(\alpha(a,p)\rightarrow\ \mathsf{A}(\delta a\rightarrow \mathsf{F}p))$ "if x does a for p, then all paths 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. $\mathsf{AG}(\iota p\rightarrow\ (\bigvee a : \alpha(a,p)))$ "if x intends p then there is an action he can do for it"

6. $\mathsf{AG}(\iota p \wedge \alpha(a,p))\rightarrow\ \beta(Pr(\mathsf{F}p|\delta a) > Pr(\mathsf{F}p))$ "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."

Semantics. The semantics of $\mathcal{R}$ is given relative to intensional models. The formal model is as described informally in section 3. Let $M = \langle \mathbf{T}, <, [\![\,]\!], \mathbf{B}, \mathbf{I}, \mathbf{A}, \Pi\rangle$ be a model. Here $\mathbf{T}$ is a set of possible moments ordered by $<$, which is finitely branching and discrete; $[\![\,]\!]$ assigns intensions to atomic propositions and actions. Π assigns probabilities to paths in $\mathbf{P}_t$, for each moment t.

The intension of an atomic proposition is the set of moments where it is true. The intension of an action is the set of periods in which an instance of it is performed. Thus, $[t,t'] \in [\![d]\!]$ means that the agent does action d from moment t to t'. For example, in Figure 1, our agent does action a in the periods $[t_0,t_1]$ and $[t_0,t_2]$. Therefore, $[t_0,t_1],[t_0,t_2] \in [\![a]\!]$. We require that each action instance has a unique termination moment. Additional coherence requirements are informally introduced as needed. It is useful to define two paths as action-equivalent if they begin with the same action (by our chosen agent). In other words, $P \approx P'$ iff $P^0 = t$ & $P'^0 = t$ & $(\forall a : (\exists t' : t' \in P$ & $[t,t'] \in [\![a]\!]) \Leftrightarrow (\exists t' : t' \in P'$ & $[t,t'] \in [\![a]\!]))$.

We define **I**, **B**, and **A** as alternative relations to give a semantics for ι, β, and α, respectively. Each of **I**, **B**, and **A** relates a path to a set of paths. An agent may have different beliefs and intentions along different paths emerging from the same moment. This enables us to treat different intentions and beliefs as options or choices available to the agent. Properties of the alternativeness relations can be used to capture various logical properties of the primitives. The main logical properties of interest to this paper are as follows:

L1. $(\forall P_B : P_B \in \mathbf{B}(P) \Rightarrow \mathbf{I}(P_B) \subseteq \mathbf{I}(P))$.

This corresponds to the constraint that ιp entails $\beta \iota p$, which means that if an agent intends something, he believes that he intends it. In other words, the agent introspects over his intentions.

L2. $(\forall P_B : P_B \in \mathbf{B}(P) \Rightarrow \mathbf{B}(P_B) \subseteq \mathbf{B}(P))$.

This corresponds to the constraint that βp entails $\beta\beta p$, which means that if an agent believes something, he believes that he believes it. In other words, the agent has positive introspection over his beliefs.

L3. $(\forall P_B : P_B \in \mathbf{B}(P) \Rightarrow \mathbf{A}(P_B) \subseteq \mathbf{A}(P))$.

This corresponds to the constraint that αp entails $\beta \alpha p$, which means that if an agent is acting for something, he believes that he is acting for it. In other words, the agent knowingly performs his actions for an intention.

L4. $\mathbf{A}(P) \subseteq \mathbf{I}(P)$.

This means that an agent who is acting for an intention intends the condition he is acting for.

L5. $\mathbf{A}(P) \neq \emptyset \Rightarrow P \in \mathbf{A}(P)$.

This means that if an agent is acting for an intention, then he is in fact performing the action on the given path. The agent may not be able to act for every intention, possibly because he lacks conviction in the suitability of an action for that intention. However, if he acts for an intention, then he does so consistently with the given path. One might wonder if the agent should be performing an action other than on the current path—in that case, $\mathbf{A}(P)$ can be empty on the given path, and nonempty on some other path. This would force his chosen action to be along one of the other paths.

Notice that we do not assume $(\forall P_A, P'_A : (P_A \in \mathbf{A}(P) \ \& \ P_A \approx P'_A) \Rightarrow P'_A \in \mathbf{A}(P))$. This would entail that the agent must choose all or none of the potential paths along which an action may be performed. Qualitative theories of know-how typically require this all or none selection [Singh, 1994], but we do not require that the agent have the necessary know-how. We develop alternative postulates in section 4.

Given his beliefs and intentions, an agent is constrained to act in a certain manner, and to add or drop additional intentions and beliefs. This is the dynamic aspect of the semantic formulation. Intuitively, it is easy to see why a static formulation, which applies to all paths from a given moment, just would not provide the expressive flexibility to model evolution and action. This is because a static formulation would yield one answer (about whether an agent has a given intention) at each moment, not an answer relativized to the path that might execute.

The semantics of formulae of $\mathcal{R}$ is given relative to a model and a moment in it. $M \models_t p$ expresses "M satisfies p at t." $M \models_P p$ expresses "M satisfies p at moment t on path P," and is needed for the "path-formulae"—which are evaluated on a path. p is *satisfiable* iff for some M and t, $M \models_t p$. p is *valid* in M iff it is satisfiable at all moments in M. 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 $M \models_t p \wedge M \models_t q$

S3. $M \models_t \neg p$ iff $M \not\models_t p$

S4. $M \models_t \mathsf{A}p$ iff $(\forall P : P \in \mathbf{P}_t \rightarrow M \models_P p)$

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. $M \models_P p\mathsf{U}q$ iff $(\exists t' : M \models_{P\downarrow t'} q \wedge (\forall t'' : t \leq t'' \leq t' \rightarrow M \models_{P\downarrow t''} p))$

S7. $M \models_P \beta p$ iff $(\forall t' : t' \in \mathbf{B}(P) \Rightarrow M \models_{t'} p)$

S8. $M \models_P \iota p$ iff $(\forall P' : P' \in \mathbf{I}(P) \Rightarrow M \models_{P'} \mathsf{F}p)$

An agent intends something if it will occur on all paths that are chosen by his alternativeness relation for intention. We insert the F operator in the evaluation to simplify the formalization, since intentions are inherently future-directed (recall that F means "now or later on the given path").

S9. $M \models_P \delta a$ iff $(\exists t' : t' \in P$ & $[t, t'] \in [\![a]\!])$

S10. $M \models_P \alpha(a,p)$ iff $(\forall P' : P' \in \mathbf{A}(P) \Rightarrow M \models_{P'} \delta a)$

S11. $M \models_P p$ iff $M \models_t p$, if semantic rules S6, S7, S8, S9, and S10 do not apply on p

S12. $M \models_t (Pr(p) = r)$ iff $[\sum_{P \in \mathrm{P}_t \wedge M \models_P p} \Pi_t(P)] = r$

The probability of p holding is r iff the probabilities of all the paths that satisfy it add up to r. (We assume that this sum exists for the propositions of interest.) Extending the syntax enables us to write and evaluate expressions involving conditional probabilities, using the definition $Pr(p|q) = Pr(p \wedge q)/Pr(q)$. As is customary, we define $Pr(p|q) = 0$ when $Pr(q) = 0$.

Notice that probabilities are evaluated at moments.

4 Postulates for Intentions

We now turn to the possible postulates, and discuss and formalize them one by one. By bringing out the undesired ramifications of the postulates, we identify more refined versions that capture the appropriate intuitions. Finally, we propose a small set of postulates as having the essential properties and avoiding the problems of section 2.

4.1 Relationship with Beliefs

P1. Intended propositions must be believed possible.

$\mathsf{A}(\iota p \rightarrow \beta(\mathsf{EF} p))$

P2. Intended propositions must not be believed inevitable. This is one of the ways in which an intention can be dropped.

$$\mathsf{A}(\iota p \rightarrow \neg\beta(\mathsf{AF}p))$$

Although these postulates do not involve actions, they are sufficient to block the inferences that cause Dudley Dolittle's problem. This is because Dudley's intention now 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. Thus Dudley cannot drop his intention to save Nell until it becomes inevitable that she will be saved. Despite this success, these postulates are still too weak:

- Many naturally arising conditions that are intended would be believed possible anyway. Thus, P1 can often be redundant.

- And, while P2 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, the conditions an agent would seriously intend might never be believed inevitable (until they finally occur).

More useful formalizations of the intuitions behind P1 are presented in P3, P5, P6, and P7. Similarly, P2 represents a useful intuition that is captured by our final set of postulates.

4.2 Acting for an Intention

It might appear that incorporating actions explicitly would help us solve Les Lazy's problem. A more fine-grained postulate is that an intention should entail that the agent holding it believes it possible that he will act for it (along some path). This postulate captures the intuition that an intention should lead to something different than if it weren't held. At the same time, this postulate allows fortuitous circumstances under which the agent might not have to act.

P3. Intentions must be believed possible to act for.

$$\mathsf{A}(\iota p \rightarrow \beta(\mathsf{EF}(\bigvee a : \alpha(a,p))))$$

On the one hand, P3 requires a possible action; on the other hand, it does not guarantee even possible success.

Lemma 1 P3 and P1 do not entail each other. ∎

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 currently adopted a plan corresponding to his intention. This postulate is apparent in the suggestion that intentions be treated as "adopted goals" [Georgeff, 1987], and is sufficiently strong to capture the relationship between intentions and actions. This would force Dudley to do something to save Nell's life.

P4. Intentions are "adopted goals." That is, they must be acted for immediately.

$$\mathsf{A}(\iota p \rightarrow (\bigvee a : \alpha(a,p)))$$

Lemma 2 shows that P4 is stronger than P3. It requires not just possible but immediate action. The result follows from the temporal logic theorem that $p \rightarrow \mathsf{EF}p$, and the assumed introspection of the agent over his actions for an intention.

Lemma 2 P4 entails P3. ∎

4.3 The Point of Intending

Les's postulate is that 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.

P5. An agent intends something only if he believes that the intended condition does not occur on paths where he does not have the intention.

$$\mathsf{A}(\iota p \rightarrow \beta(\mathsf{A}(\neg\iota p \rightarrow \neg\mathsf{F}p)))$$

An alternative formalization might use $\neg\mathsf{EF}p$ instead of $\neg\mathsf{F}p$ in the above formula. This would fail, because $\neg\mathsf{EF}p$ is evaluated at moments, rather than along paths. If $\neg\mathsf{EF}p$ is true at a moment, it is true even along paths from that moment on which the agent does not intend p.

P5 requires that the agent believe that his intention is necessary for p. This makes sense only for conditions that are fully under the control of the agent and cannot be caused by others. (About the only such conditions are an agent's speech acts [Austin, 1962, p. 7] or conditions involving an agent's cognitive states.) However, most conditions 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.

Consequently, P5 is too strong. By contrast, P6 states that while the intended condition might occur even on paths where the agent does not have the given intention, it cannot be guaranteed to occur on those paths. On such paths 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.

P6. An agent intends something only if he believes that the intended condition is not inevitable on paths where he does not have the intention.

$\mathsf{A}(\iota p \rightarrow \beta(\mathsf{A}(\neg\iota p \rightarrow \neg\mathsf{FAF}p)))$

Lemma 3 P5 is strictly stronger than P6. ∎

Lemma 3 follows from the properties of the temporal conditions involved in the two postulates. Briefly, $\neg\mathsf{F}p$ entails $\neg\mathsf{FAF}p$, but not vice versa. Since P6 is the weaker of the two, it might more easily acceptable. However, both postulates can each lead to trouble! This is because they validate the following inference, which is nothing but Les Lazy's problem, as described in section 2.2.

- $\mathsf{A}(\iota p \wedge \beta(\mathsf{AG}(q \rightarrow p)) \rightarrow \neg\iota q)$

  or put slightly differently:

- $\mathsf{A}(\iota q \wedge \beta(\mathsf{AG}(q \rightarrow p)) \rightarrow \neg\iota p)$

In other words, if an agent intends q and believes that q entails p in the future, then he cannot intend p. This inference might be termed *consequential anti-closure:* a most unusual problem for a logic of cognitive concepts!

4.4 Commitments

Instead of trying to work around consequential anti-closure, we formalize the above requirement using *commitments* [Bratman, 1987, ch. 2] [Singh, 1996]. An agent has an intention only as long as he is committed to it. This notion of commitment is internal or mental, and must be carefully distinguished from social commitment, which is also discussed in the AI literature [Castelfranchi, 1995; Singh, 1998]. As before, let q entail

p. Intuitively, if the agent acts for q on the basis of his reasons for q, he may not only achieve q, but also achieve p as a side-effect (in fact, he might achieve p even if cannot achieve q). Suppose the agent fails in achieving q, does not achieve p, 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. The commitments for both p and q are held concurrently.

A natural way of treating commitments formally is by using probabilities. Each commitment should improve 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 intuition behind P5 may be expressed as in P7 below. Consequently, P7 replaces P5 and P6.

P7. An agent believes that his intention raises the probability of success.

$$\mathsf{A}(\iota p \rightarrow \ \beta(Pr(\mathsf{F}p|\iota p) - Pr(\mathsf{F}p|\neg\iota p) > \epsilon))$$

Here $0 < \epsilon \leq 1$ is a parameter telling us how much likelier is likelier enough. P7 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.

P7 is perhaps not obvious. We won't defend it strenuously here, because we shall replace it with another postulate shortly. However, we should point out that because probabilities are evaluated at moments, P7 is neither inconsistent nor trivial, as one might initially suspect. For example, even if $P \in \mathbf{I}(P)$, i.e., $\mathbf{I}$ is reflexive or the agent's intentions are realized in the given path, $\mathsf{F}p$ would hold on the given path, but not necessarily on every path that is generated from the initial moment of the given path.

But in its present form, P7 does not help 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. The key idea is that although an agent is committed for acting for his intentions, he is not committed to doing so now, only *eventually*.

P8. An agent with an intention will act for it eventually, if the intention does not come to hold on its own.

$$\mathsf{A}(\iota p \rightarrow \mathsf{A}(\mathsf{F}p \vee \mathsf{F}(\bigvee a : \alpha(a,p))))$$

Lemma 4 P4 is strictly stronger than P8. ∎

Lemma 4 follows from the entailment relationship among the given temporal logic formulae. Because of Lemma 4, the above postulate replaces P4, which is one of the sources of Ken's mistake.

P9. An agent acts for an intention only when he believes that doing so would significantly raise the probability of success. (Recall that $\alpha(a,p)$ entails ιp.)

$$\mathsf{A}(\bigwedge a : \alpha(a,p) \rightarrow \beta(Pr(\mathsf{F}p|\alpha(a,p)) - Pr(\mathsf{F}p|\neg\alpha(a,p)) > \epsilon))$$

P9 replaces P7, which is the other source of Ken's mistake. This postulate captures the intuitions that it is not just having an intention, but acting for it that should raise the probability of success. That is, the agent should act only if he needs to, and as we can see from P8, an agent might not always need to act for his intentions.

4.5 Solution

Because of the above reasons, our solution must include P8 and P9. Lemma 5 shows that P3 combined with P8 and P9 entails P2. This is because P3 requires that the agent believe he will act for his intention. 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 it could not have been believed inevitable.

Lemma 5 P9, P8, and P3 jointly entail P2. ∎

Therefore, we adopt P8, P9, and P3. Again, since $\epsilon > 0$, P9 entails that, if the agent intends p (and acts for it), he believes that the probability of its occurrence is nonzero, i.e., it will occur on some path. Thus, given P3, P9 entails P1. Further, Lemma 6 establishes that the three constraints can all be satisfied together. Thus we have nonparadoxical formulations of the intuitions behind the postulates on intentions.

Lemma 6 P9, P8, and P3 are jointly consistent. ∎

Hence, we obtain the following result. This is an informal claim, but this paper has sought to motivate it, formalize it in a manner that preserves our intuitions, and prove it correct.

Theorem 7 P9, P8, and P3 solve our problems. ∎

An unexpected consequence of P3 and P4 is that if an agent has an intention, he must believe that at some possible future his acting for it will improve his chances of success. Suppose in no situation can an action chosen by the agent improve the believed likelihood of success. Then the agent should simply plan to perform any action other that what his initial plan (his "instincts" if you like) suggest. If some such α proves appropriate, we are fine. If no such α exists, then the agent must be unable to improve the likelihood of success at all. However, this situation violates our intuitive assumption that intentions are intimately related to actions.

5 Discussion

This paper presented a rich framework for representing intentions, actions, and beliefs that involves probability and choice. This framework uses ideas from previous temporal models, but can express some intuitive, but troublesome, rationality postulates on intentions. Unlike traditional logical approaches, we treat intentions and beliefs as path formulae. This might appear less intuitive, but we believe that it is essential to be able to perform hypothetical reasoning about intentions and beliefs. In discrete models, one might encode paths in successor moments, but this would fail for nondiscrete models.

Some of the present theories of intentions have some components of dynamism. For example, [Cohen and Levesque, 1990] require that intentions be persistent, and embody this requirement in the semantics of intentions. A number of authors, e.g., [Rao and Georgeff, 1991; Singh, 1994], have pointed this out as inappropriate, because it is both (a) rigid and (b) mixes the definitions of intentions with reasoning about intentions. [Singh, 1994] develops a qualitative framework in which constraints on intentions, such as persistence, can be expressed. This work allows nondeterministic models in which it formalizes know-how. It shows how an agent with the right know-how can succeed with an intention if he selects actions from among those that will lead to success [Newell, 1982]. [Georgeff and Rao, 1995] formalize some properties of intention maintenance in a qualitative framework. They assume full determinism, and do not consider actions. However, they consider the interesting special case where an agent has only one intention in which some inferences can be drawn qualitatively, and without considering actions. [van Linder *et al.*, 1996] formalize various attitudes including commitments. They capture a definition of "commit-to" that recalls the action selection condition of [Newell, 1982]. Although the above approaches are useful contributions, in lacking a quantitative and explicitly normative stance, they fail to address crucial components of intention dynamics. As we argued, qualitative approaches cannot accommodate all of the three intuitive problems described in section 2.

[Wobcke, 1995] studies plans and intention revision in a formal framework. Wobcke independently observes the static nature of traditional approaches to the semantics of intentions. However, he does not propose a path-based semantics for intentions as here. Wobcke uses ideas from belief revision, including the rationality postulates for belief revision due to [Gärdenfors, 1988]. However, he does not propose any postulates involving intentions, as we have attempted to do. Instead, Wobcke offers only specific examples of planning that involve adopting additional intentions based on current intentions and beliefs.

[Singh, 1996] develops a utility-based approach to commitments and precommitments. Singh attempts to formalize the conditions under which an agent should deliberate and the "conative" policies that affect when and how that deliberation is carried out. The present approach addresses the key postulates on intentions that avoid the problems discussed above; these postulates by themselves can be captured using only probabilities instead of utilities—in fact, a simpler approach helps highlight the intuitions involved. However, we believe that, if necessary, 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 giving a more concrete account of how limited rational agents adopt and drop their intentions.

This paper has only hinted at the kinds of postulates that would need to be captured in order to obtain a sufficiently complete theory of the dynamics of BDI agents. We have sought to address the chief conceptual problems such as the dynamism of the semantic definitions and the subtle relationships between objective probability and the internal states of agents. However, the space of postulates still remains to be explored in depth. This is an important future direction.

References

[Austin, 1962] John L. Austin. *How to Do Things with Words.* Clarendon Press, Oxford, 1962.

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

[Breiter and Sadek, 1996] P. Breiter and M. D. Sadek. A rational agent as a kernel of a cooperative dialogue system: Implementing a logical theory of interaction. In *ECAI-96 Workshop on Agent Theories, Architectures, and Languages*, pages 261–276. Springer-Verlag, 1996.

[Castelfranchi, 1995] Cristiano Castelfranchi. Commitments: From individual intentions to groups and organizations. In *Proceedings of the International Conference on Multiagent Systems*, pages 41–48, 1995.

[Chellas, 1992] Brian F. Chellas. Time and modality in the logic of agency. *Studia Logica*, 51(3/4):485–517, 1992.

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

[Doyle, 1992] Jon Doyle. The roles of rationality in reasoning. *Computational Intelligence*, 8(2):326–335, May 1992.

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

[Gärdenfors, 1988] Peter Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, Cambridge, MA, 1988.

[Georgeff and Rao, 1995] Michael P. Georgeff and Anand S. Rao. The semantics of intention maintenance for rational agents. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 704–710, 1995.

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

[Grosz and Sidner, 1990] Barbara Grosz and Candace Sidner. Plans for discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *SDF Benchmark Series: Intentions in Communication*, pages 417–444. MIT Press, Cambridge, MA, 1990.

[Haddawy, 1996] Peter Haddawy. Believing change and changing belief. *IEEE Transactions on Systems, Man, and Cybernetics Special Issue on Higher-Order Uncertainty*, 26(5), May 1996.

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

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

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

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

[Pollack, 1992] Martha E. Pollack. The uses of plans. *Artificial Intelligence*, 57(1):43–68, 1992. Computers and Thought Award Lecture.

[Rao and Georgeff, 1991] Anand S. Rao and Michael P. Georgeff. Asymmetry thesis and side-effect problems in linear-time and branching-time intention logics. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 498–504, 1991.

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

[Singh, 1996] Munindar P. Singh. Commitments in the architecture of a limited, rational agent. In *Proceedings of the Workshop on Theoretical and Practical Foundations of Intelligent Agents*, pages 72–87. Springer-Verlag, 1996.

[Singh, 1998] Munindar P. Singh. An ontology for commitments in multiagent systems: Toward a unification of normative concepts. *Artificial Intelligence and Law*, 1998. In press.

[van Fraassen, 1981] Bas C. van Fraassen. A temporal framework for conditionals and chance. In *[Harper* et al.*, 1981]*, pages 323–340. 1981.

[van Linder *et al.*, 1996] B. van Linder, W. van der Hoek, and J.-J. Ch. Meyer. Formalising motivational attitudes of agents: On preferences, goals and commitments. In *Intelligent Agents II: Agent Theories, Architectures, and Languages*, pages 17–32, 1996.

[Wobcke, 1995] Wayne Wobcke. Plans and the revision of intentions. In *Proceedings of the Australian Workshop on Distributed Artificial Intelligence, LNAI 1087*, pages 100–114, Heidelberg, 1995. Springer-Verlag.