Intentions Are Plans Plus Wishes (And More)

Roberto Zamparelli
Dept. of Foreign Languages,
Literature and Linguistics
University of Rochester
roberto@psych.rochester.edu

Abstract

Cohen and Levesque ([CoL90a], [CoL90b]) propose a logic of "rational actions" and give a possible-worlds model interpretation to the propositional attitudes of *belief* and *goals*, defining from these the notion of *intention*. In this paper, I will discuss various problems with this accont and propose augmenting their logic with modality, two predicates expressing the notions of "interest" (WISH) and "commitment to an action" (COMMITMENT), and a weak version of "plan"(D-PLAN). The representation of "intentions" will be based upon a general theory of the relation between beliefs, desires and intentions.

Introduction

A satisfactory representation of human intentions requires an understanding of the relationship between the *cognitive* component of the mind, its *motivational* component, and the actions that results from the interaction of the two. In philosophy of mind much has been written on cognition and action (see [Sea83], [Bra87], [Bra87], [Cas75]), less on the motivational component of an agent. *Why* should an agent, given certain information, decide to act on its environment? Applied to a computer program, the question has a straightforward answer: its programmer has wired it to do so. The motivation of the programmer was transferred to the program, in the form of instructions in a procedural programming language. Later on, as A.I. attempted to model an *independent, planning* agent, the motivational aspect was transferred into predicates such as "GOAL" and "PLAN". An agent acts *because* it has a GOAL to achieve, or a PLAN to follow a specific course of actions.

Much confusion arose, however, as people failed to agree on a common, intuitive meaning for these predicates. GOAL, for instance, could render the expression "active goal" (i.e. a state of the world that an agent has started acting toward), "background goal", a disposition, or inclination to act, more related to the English noun "desire", or even "consequence of one's realistic goals" [CoL90]; similarly, PLAN could be a mere "recipe for acting", or a committed will to act along pre-specified lines.[1]

From a linguistic standpoint, similar problems arise from the attempt to map directly such "vague" English terms as "plan", "goal", "want", "intention", etc. into unambiguous first-order-logic predicates, when semantic analysis might reveal a genuine ambiguity which must be torn apart before formalization is possible.

In [Zam9?] I argue that the meaning of attitude predicates in a natural language is both vague and ambiguous. In particular, I maintain that the distinction of between "plan-as-recipe" and "active-plan", or that between "background goal" and "active goal" is in the notion of *commitment*.

"Commitment" can be characterized as the state which is the result of a decision to act. I will take it as a primitive notion, and suggest that it is essential in the representation of the mental states of *intentions* and *default beliefs*. But first, it must be presented within an analytical theory of mind which I try to spell out in some detail in [Zam9?].

The basic idea is that there are two kinds of mental attitudes, *cognitive attitudes* and *motivational attitudes*.[2] For ease of discussion, I will use *beliefs* and *desires* as representative members for the two classes.[3] Both denote *states*, and both have as content *states of affairs*, (*propositions*, sets of (structured) possible worlds, or any other suitable representation). The most important distinction between the two is that beliefs *per se* are completely inert data, while desires characteristically motivate actions.

In a computational metaphors of mind, cognitive attitudes are PROLOG-style clauses, while motivational attitudes are instructions in a procedural programming language; when executed, the latter cause the com-

[1] See [All90]. The difference between plan-as-recipe and active-plan is captured by the difference in English between the expressions "I plan to X" (an intention) vs. "I have a plan to X" (mostly "recipe").

[2] In this work, I will completely set aside the important role of predicates of *emotion* and *perception*.

[3] For a useful synopsis of propositional verbs see [Ven72].

puter to act in certain ways, perform some actions, examine data, etc.

Together, cognitive and motivational attitudes cause an agent's actions (see [Brn87]). The action, then, can be physical or mental; a *decision* (together with an *attempt*, a *choice*, etc.) is an example of mental action, and a *commitment* is, together with an immediate physical action, the possible result of a decision. If we assume that actions to be executed are put in a TO-BE-DONE buffer, expressing a commitment means reporting part of the content of this buffer. Again, in the computer metaphors, a *decision* is the action performed by a running program when it executes a conditional instruction.

Notice, as an important difference between beliefs/desires and decisions that only the former can be *graded*. One cannot decide "a bit", or "partially". Of course, he can take a decision based on weak motivations, which means that he will be inclined to reconsider it. But reconsidering has nothing to do with the growing and fading of a desire, or a doubt.

2. Intentions and Default Beliefs

Bratman [Bra87] discusses at length the main reason why a rational agent with limited resources should have a notion of "intention". Without intentions, the agent might have to compute endlessly what her (or worse, other's) future course of actions might be, in the face of a rapidly changing environment. With intentions, on the other hand, she can assume that, by default, the intended action *will* be done and start working from this. Intentions are the result of a decision, not a stage of the process of deciding; as such, they are premises to further planning *as if* they stood for the action.

As a working hypothesis, I will suggest that keeping a possible action (or a possible belief) in the consequent of an open conditional requires a greater computational effort than just assuming that it will be done (or accepted), despite the fact that at times this strategy will go wrong. This hypothesis is not purely speculative; in fact, it is supported by recent psychological studies on the way people parse sentences (see [Gib90]). It appears that, faced with the lexical and syntactic ambiguities present in any sentence, people do not process all the possible parses in parallel, but establish a limited commitment to the one parse that seems to be favored by lexical, syntactic and semantic factors. Processing in parallel many possible paths amounts to keeping in working memory a number of nested conditionals, to be pruned as the sentence proceeds. However, this seems to happen to a very limited extent (see [WeiIP]). The case of future actions and beliefs is, I suggest, analogous, and both point to computational limits.

If this is correct, intentions, once formed, are unconditioned. If we "change" an intention, it is because the change of some external or internal factor prompted us to redo the whole decision process, and this returned a different intention. In other terms, intentions can only be *replaced*, never *modified*.[4]

Anything said so far about intentions should really be said of their essential ingredient, *commitment*. I believe that intentions must be analyzed as a commitment to perform an action plus a plan to do so.[5] However, the plan component is only responsible for the fact that intentions are realistic (see section 8).

The same functional motivation given for intentions can be applied to another component of mental machinery, namely, *default beliefs*. The problem of default beliefs is to establish what kinds of conclusions can be derived from a set of premises via non-monothonic inference rules. Adopting the closed-world assumption, the hard question is to decide, in case of conflicting defaults, which one will prevail (see [Gef90]). In addition, a realistic autonomous agent in an open world will have to solve this problem with limited resources, i.e. without the possibility to exhaustively search its knowledge base for possible counterexamples. Given our working hypothesis about computational load (it is more expensive to keep a possible belief inside a conditional that to assume it, on risk of being wrong) the agent will find convenient to entertain a hypothesis as a provisionary conclusion, unless indication to the contrary has emerged within a certain time (or effort).[6]

Doing this amounts to *deciding to believe*, i.e. establishing a limited commitment to assume a particular piece of knowledge.[7] Notice the functional similarities between default beliefs and intentions. Both are heuristic strategies designed to ease the computational load of an agent, both are clearly defisible. From a structural point of view, a formal predicate for *commitment* will take actions as arguments; as a consequence, beliefs should have the form of *maintainance actions*,

[4] It is this intuition about intentions what makes me most suspicious of the possibility to define a predicate INTEND using *relativized* goals (e.g. "P-R-GOAL" in [CoL90a]).

[5] Although intentions are *caused by* a motivational state (WISH, as defined in section 5), since I submitted the title of this paper I have had doubts about whether WISH is actually an *analytical* component of an intention. The question boils down to the internal partitions for the accessibility relation WISH needed to distinguish (uncommitted) "desiring" to (uncommitted) "wanting". So, while one can imagine a situation in which an agent A *intends to X but doesn't desire to X* it is doubtful whether the same situation can be described as *A intends to X but doesn't want to X*. I discuss this example in Zamparelli (forthcoming).

[6] In this case, the problem of defaults becomes to introspect knowledge beginning with those portions that are more likely to be relevant for the truth value of the default conclusions.

[7] A "decision to believe" is reminiscent of the use of preference for sorting out competing beliefs in [Gal89], although on the appropriateness of Gallier's predicate PREFER see section 6.

resulting from an action of "assuming", "learning", etc. The relation between the state of belief and the action that originated should be analogous to the relationship between the state of commitment in an action and the corresponding decision, choice, etc.

However, in the rest of this paper I will not pursue this line of analysis, and I will stick to a traditional, possible-world representation of beliefs, in order to focus on various features of intentions.

3. Semantics and Pragmatics

I want to suggest that the ambiguity in various attitude verbs in English is due to presence of lack of the predicate expressing commitment in their semantics. Intentions are a combination of a commitment by the agent to do action A and a plan to do A. It follows from this that one can only "intend to do an action" (indeed, as we shall see, a *voluntary* action), never "intend a state". Before discussing possible objections and alternative treatments of intentions in the next sections, I should give reasons why, beside commitment, an intention also requires a plan. This is to distinguish "intention" from one of the meaning of the verb "want", i.e. *committed-want.* In my intuitions, there is a meaning of "want" which express more than mere desire (it hides a real will to act toward the goal) but which is not interchangable with "intend", since "intend" requires some idea of how to bring about the desired state, while *committed-want* doesn't.[8] Committed-want then, would be analyzed as commitment without necessarily a plan do go with it.

A first possible objection to the view that an intention is the state which follows a decision to act is that sometimes we seem to have intentions without the impression to have decided anything. Decisions, it seems, require effort, thought, sometime a full-night sleep; but we can find ourselves having the intention to do something without even realizing how it arose. This objection, however, can be easily dismissed. It is not "deciding" in itself that we tend to notice, but those decisions that require a choice between two possibilities of comparable value. In most cases, we don't notice taking a decision because there is no real competition among the various possibilities. One wins immediately, or the contrast is of such minor importance that can be resolved randomly.

A more serious objection is linguistic. If intending takes only voluntary actions and not states, what about sentences like (1–2)?

(1) "I intend to be rich"
(2) "I intend that you leave this room immediately"
(1′) ??"I intend that the sun is up tomorrow"
(2′) ??"I intend that this mosquito goes to bite my roommate"

It must be noted here that, first, the possibility of introducing states or alien actions after "intend" is very restricted (witness (1′–2′)). When we do not seem to have any control on the way to achieve the intended state, the sentence becomes worse. Similarly, (2) suggests that the speaker has some kind of power on the hearer; when this power is lacking, as in (2′) the sentence becomes odd. Second, there is a great deal of variability among speakers as to the acceptability of (1) and (2).

All these facts point to a pragmatic explanation of cases (1–2). At a semantic level, the predicate *intend* takes only actions; however, some speakers give it leeway to take states *when it is pragmatically clear what actions the speaker could do to achieve the desired state.*[9] (2) can be explained via conversational implicatures [Gri75]; by uttering (2) the speaker S wants the hearer H to come to believe that S has upon H the same power that S has on his own voluntary actions; in particular, that H can be used by S as part of S's plan to send H away, and that, therefore, S has the intention to send H away using H's comprehension of the situation and knowledge of S's power on him.

An alternative analysis of intentions tries to reduce them to a combination of believes and desires. Bratman [Bra87] and Brand [Brn87] argue at length against this idea, pointing out that one can have the belief that he will do something, the desire to do it, but not being really "intending" to do it.[10] In the following sections I will analyze a related reductionist definition of "intention" given in [CoL90a], [CoL90b]. Here two predicates for intention are reduced to two primitive predicates, GOAL and BEL.[11]

[8] Thanks to Rosaria Conte for pointing out to me the importance of committed-want, as exemplified, for instance, by "I want to win!"—said by an athlete before a race—or "I want there to be peace and freedom for every man!", said by a social activist. These can be much more that wishes, but hardly intentions.

[9] This is analogous to the case of verbs like "want" which can take both sentences and NP's ("I want ice-cream" "I want to swim"). A possible analysis of these cases (Schubert, p.c.) is that "want" always takes a sentence, but that in certain cases the verb must be pragmatically filled-in. So, "I want ice-cream" = "I want to {have / eat / ...} ice-cream

[10] For instance, suppose that some friends are pressing me to go to the cinema. I would like to go to the cinema but right now I don't intend to, although I am practically sure that they will end up convincing me and that I will eventually go. The contrary is more problematic, see fn. 5.

[11] Where confusion cannot arise, I will use the term "(P-)GOAL" (and "(P-)GOALs") to refer ambiguously to the predicate and to the state(s) P of which (GOAL x P) is true. Same for WISH(es) below.

4. Cohen & Levesque [CoL90b] system

C&L [CoL90b] logic (hereafter $L_{C\&L}$) is built upon four primary operators, BELief, GOAL, HAPPENS and DONE. BEL and GOAL are in turn based on $\mathcal{B}$ and $\mathcal{G}$, accessibility relations on worlds. Worlds (or *situations*[12] are modeled as *courses* (i.e. *sequences*) *of primitive events* . Informally, $\mathcal{B}$ defines the set of worlds which an agent regards as possible (or *open*) at a given time/world, and $\mathcal{G}$ the set of worlds which an agent believes would be true if his realistic goals were satisfied. Worlds in $\mathcal{G}$ include both states that the agents believes to be already true and states that are now believed possible and that the agent believes would then become true. (BEL x P), "x believes that proposition P is true", expresses the fact that all worlds accessible to x via $\mathcal{B}$ (hereafter $\mathcal{B}_x$) contain P, while "believing that proposition P is possible" (or "open"), is expressed by saying that P is present in some (but not all) worlds in $\mathcal{B}_x$. Correspondingly, (GOAL x P) "P is an expected consequence of the fulfillment of x's goals" (i.e. P is part of what x believes would be true if his goals were fulfilled) means that P is present in all the worlds in $\mathcal{G}_x$.[13]

HAPPENS and DONE express respectively the idea that some event is going to happen *immediately after* the evaluation time, or has happened *right before* the evaluation time. Since coincident events are excluded, a world w and an integer n are sufficient to identify univocally the nth event in world w.

C&L propose two important constraints on $\mathcal{B}$ and $\mathcal{G}$. They are:

- *Consistency*: $\mathcal{B}$ is Euclidean, transitive and serial, $\mathcal{G}$ is only serial.[14] The *necessitation rule* is adopted: if $\models \alpha$ then $\models$ (BEL x α).
- *Realism*: $\forall\sigma$, σ^* , if $< \sigma$, n$> \mathcal{G}$[p]σ^*, then $< \sigma$, n$> \mathcal{B}$[p]σ^*. In other terms, $\mathcal{G} \subseteq \mathcal{B}$.

[12] C&L leave open whether the function which associates integers with events is a total function (i.e. if worlds are temporally infinite), or whether *events*, as changes from states to other states, require for the states to be represented *globally*, i.e. to be a conjunction of all facts.

[13] "GOAL" is arguably a misnomer; see [All90].

[14] For $\mathcal{B}$ to be *transitive* means that whenever (s, u) $\in \mathcal{B}$ and (u, t) $\in \mathcal{B}$ also (s, t) $\in \mathcal{B}$; to be *Euclidean*, that whenever (s, t) $\in \mathcal{B}$ and (s, u) $\in \mathcal{B}$, also (t, u) $\in \mathcal{B}$; to be *serial*, that for every world s there is a world u such that (s, u) $\in \mathcal{B}$. The fact that $\mathcal{B}$ is transitive, Euclidean and serial ensures the soundness of, respectively, $\forall$x (BEL x P) $\rightarrow$ (BEL x (BEL x P)), $\forall$x $\neg$(BEL x P) $\rightarrow$ (BEL x $\neg$(BEL x P)) and $\forall$x (BEL x P) $\rightarrow$ $\neg$(BEL x $\neg$P), which, together with $\forall$x ((BEL x P) $\wedge$ (BEL x (P $\rightarrow$ Q))) $\rightarrow$ (BEL x Q), are the fundamental axioms for weak S5 modal logic (see [HaM85], [FaH85]. Given the "actual" world i, this properties ensure that the set {t | (i, t) $\in \mathcal{B}$ } is reflexive (i.e. the actual world is believed to be possible), symmetric and transitive. Both $\mathcal{B}$ and $\mathcal{G}$ are serial, which insures that for each world s there is always some world accessible from s via these relations.

Given our definition of (BEL x P) (P is true in all the worlds accessible to x via $\mathcal{B}$), we have that the following hold:

1. $\models \forall$x(BEL x P) $\rightarrow$ (GOAL x P)
2. $\models \forall$x(GOAL x P) $\rightarrow$ $\neg$(BEL x $\neg$P)
3. $\models \forall$x((GOAL x P) $\wedge$ (BEL x (P $\rightarrow$ Q))) $\rightarrow$ (GOAL x Q)
4. $\not\models \forall$x(GOAL x P) $\rightarrow$ (BEL x P)

For lack of space, the reader is referred to [CoL90a, Col90b] for the formal specification of the logic. For ease of exposition, we repeat some abbreviations there defined.

$\Diamond$**A** := $\exists$**y(HAPPENS y;A?)**, *"eventually* A": A will happen after some future event.

$\Box$**P** := $\neg\Diamond\neg$**P**, *"always* P": P is always true in a given world.

(LATER P) := $\neg$**P** $\wedge$ $\Diamond$**P**

(HAPPENS x A) := **(HAPPENS A)** $\wedge$ **(AGT x A)**, where (AGT x A) is the unique agent of action A; a version of HAPPENS which specifies the agent of the future action.

(DONE x A) := **(DONE A)** $\wedge$ **(AGT x A)** where (AGT x A) is the unique agent of action A; a version of DONE which specifies the agent of the past action.

(A-GOAL x P) := (GOAL x (LATER P)) $\wedge$ (BEL x $\neg$P)

(P-GOAL x P) := (GOAL x (LATER P)) $\wedge$ (BEL x $\neg$P) $\wedge$ [BEFORE ((BEL x P) $\vee$ (BEL x $\Box\neg$P)) $\neg$(GOAL x (LATER P))

(P-R-GOAL x P Q) := (GOAL x (LATER P)) $\wedge$ (BEL x $\neg$P) $\wedge$ [BEFORE ((BEL x P) $\vee$ (BEL x $\Box\neg$P)$\vee$(BEL x $\neg$Q)) $\neg$(GOAL x (LATER P))

From these notions of "goal" C&L move on to define the two predicates for "intention", INTEND_1 and INTEND_2. The first one applies to actions ("I intend to go out"), the second one, to states which one wants to become true "without having any idea of how to achieve [them]"[15] (an example would be "I intend to become rich").[16]

(INTEND_1 x A) := (P-GOAL x [DONE x (BEL x (HAPPENS a))?;a]), where A is any action expression.

(INTEND_2 x P) := (P-GOAL x $\exists$e(DONE x [(BEL x $\exists$ e'(HAPPENS x e';P?)) $\wedge$ $\neg$(GOAL x $\neg$(HAPPENS x e;P?))]?;e;P?))

[15] [CoL90b,247]

[16] For both intention predicates C&L also give a "relativized" version which substitutes P-R-GOAL for P-GOAL. The difference is that with the new version I can INTEND something relatively to a condition Q which, if proved false, makes me drop the intention. See fn. 4.

$INTEND_1$ can be paraphrased as: an agent x has the intention to do a certain action A if he has the persistent goals that he brings about A *and* that immediately before A becomes true, x knows that the next thing to happen is that A becomes true. This second part is designed to rule out accidental accomplishment of intended action.[17] $INTEND_2$ is more complex, involving the notion of a "plan" e which the agent thinks must precede the desired state P. The plan might not be completely worked out, but there is at least one action e′ which x must believe he is going to do next and which will precede the desired state P; e′ can thus be seen as a "preliminary" step for e.

5. The folk psychology of Intentions & Co.

It is useful at this point to review some desiderata for a folk-psychology notion of "intention" and related attitudes, as presented in [Bra87], [Se83] and [Brn84], in comparison with the logical properties of the predicates in [CoL90a, CoL90b].

1. *Belief, desires and intentions are distinct mental attitudes and are not reducible to one another.* C&L express intentions as a combination of P-GOAL (a goal) and BEL (the belief that the action is going to happen, immediately before it happens), plus a notion of agenthood. Arguably, this shifts some of the less-understood aspects of the notion of "intention" onto the notion of "agenthood"; for instance, the difference in meaning between [DONE (BEL x P) ∧ (AGT x (BEL x P))] and [DONE (BEL x P)] is rather obscure.

2. *The content of an intention is an action.* I have argued that the core meaning of "intend" is incompatible with "intending a state". This goes against the need to define $INTEND_2$, which should fall off from a proper definition of INTEND (C&L's $INTEND_1$) plus an appropriate account of conversational implicatures.

3. *We can intend to achieve a certain result without intending to achieve all its expected consequences.*

4. *The expected consequences of a voluntary action are chosen.* Notice that C&L GOAL cannot model the notion of *chosen consequence*, since anything I believe to be currently true is a GOAL. In addition, as pointed out in [Sad92], everything that is believed to be currently possible but not true is automatically a GOAL. This is because (P-GOAL x P) is spelled out as (GOAL x ◇P ∧ (BEL x ¬P) ∧ [BEFORE ((BEL x P) ∨ (BEL x □¬P)) ¬(GOAL x (LATER P)). But this is realized any time we have: (BEL x (◇P ∧ ¬P)) ∧ [BEFORE ((BEL x P) ∨ (BEL x □¬P)) ¬(GOAL x (LATER P))].

[17] The rationale is ruling out a traditional puzzle, due to Crisholm, in which an agent who intends to kill his uncle accidentally runs over a pedestrian who turns out to be the uncle. Here it cannot be claimed that the agent killed the uncle "following his intentions".

5. *Intentions (unlike desires) cannot be inconsistent.* This follows naturally if we adopt the idea that intentions contain a plan component, since we cannot craft a single plan-as-recipe to achieve both A and ¬A at the same time.

6. *Intentions are believed (at least) possible.* Again, if intentions contain a plan, this follows for any conceivable definition of what a plan ought to be.

7. *People have non-realistic attitudes.* By "non-realistic attitudes" I mean that people imagine contrary-to-fact situations and desire states which they believe to be impossible. In the A.I. tradition, these phenomena are simply filtered away, under the assumption that, whatever their significance, they will not affect the agent's *rational* actions.

I believe that there are reasons to support non-realistic attitudes within a general theory of "actions", even when it is computationally motivated.

First, from a theory-internal standpoint, GOAL is not a primitive atomic predicate, as in C&L intentions, because it is based on $\mathcal{G}$, which is constrained by $\mathcal{B}$. The relation $\mathcal{G}$ can in fact be defined as the intersection of $\mathcal{B}$ and an accessibility relation expressing the fundamental notion of *positive motivation*, or *interest*, i.e. the set of situations that an agent, totally ignorant of the state of the world, *would wish were true.* I shall call such a relation $\mathcal{W}$, and the corresponding predicate WISH, informally defining (WISH x P) as "x has an interest that P is true", regardless of whether he believes P to be possible or not, and regardless of the reasons why he wants P to be true.[18] Second, beliefs can change. If we just dropped any desire that we currently believe unachievable, every time our beliefs change we should go back and see, for all our abandoned desires, which ones have become "possible" in the new circumstances. This leads rapidly to a "desire-version" of the frame problem, and—I believe—is a strong motivation to keep unachievable desires somewhere in the background.[19]

Following these intuitions, we can define a predicate for WISHes as follows. Let $\mathcal{W}$ be a structure T × P × Z × T in the logic L_{BW}, augmented from $L_{C\&L}$. Like $\mathcal{G}$, $\mathcal{W}$ will be only serial, to insure that we do not have for WISHes anything correspondent to positive and negative introspection. Now we have:

$M,\sigma,v,n \models (\text{WISH x } \alpha)$ iff for all σ^* such that $< \sigma, n > \mathcal{W}[v(x)]\sigma^*$, $M,\sigma^*,v,n \models \alpha$.

That is, "α follows from x's wishes" in world σ at

[18] WISH is a component of situations as differentiated as "craving for an ice-cream" "desiring to be a bachelor", "being willing to betray" "be willing to pay a ransom to rescue someone", and many others. The difference between, say, the first and the last of these cases is to be found in the absence vs. presence of conflicting desires. See fn. 5, [Zam9?].

[19] But compare to the role of the "Opportunity Analyzer" in [BrI88]) general architecture.

time n iff α is true in all worlds σ^* accessible to x in world σ at time n via the relation $\mathcal{W}$. Notice that, critically, $\not\models$ (BEL x P) $\rightarrow$ (WISH x P), because $\mathcal{W}$ might include worlds outside $\mathcal{B}$.

6. Expressive Power.

The peculiar aspect of $L_{C\&L}$ is that C&L use a modal semantics without modal operators. The operators $\Box$ and $\Diamond$ mean, respectively, "always" and "eventually", while the more traditional notions of "necessity" and "possibility" remain only in the specification of the predicate BEL. However, there are things that seem to inherently require some notion of possibility.

Consider *atemporal goals*, like those expressed by "I want to be happy.", "I want to become rich.", "I need water". It appears that typically we want these goals to be fulfilled *as soon as possible*, modulo the possibility to minimize the contrast with interacting conflicting goals[20]. While temporal goals can be expressed in C&L framework, expressing the notion *as soon as possible* is problematic.

Consider the following example. I hate John Barleycorn and I badly desire his death. This can be represented as a (fanatic) persistent goal that John Barleycorn is dead, i.e. (P-GOAL me Dead(jb)). This unfolds as (GOAL me ($\neg$Dead(jb) $\wedge$ $\exists$e(HAPPENS e;Dead(jb)?))) $\wedge$ (BEL me $\neg$Dead(jb)) $\wedge$ ... However, since I believe that John is currently alive and that, being a man, he is mortal, I also have that: (BEL me $\neg$Dead(jb)) and (BEL me $\exists$e$'$(HAPPENS e$'$;Dead(jb)?)), where intuitively e$'$ is "the event of John's life, however long it might be". The result is that my goal is not an active goal; I want that there is a time after which John is dead, and I believe that, indeed, this will be the case. So, I am not motivated to act.[21]

Turning the atemporal goal into a timed goal will not do. A better alternative is using a notion of "preference":

(1)
(PREFER x (HAPPENS Dead(jb)) $\exists$e(HAPPENS e;Dead(jb)))

I.e., the agent prefers that John's demise happens as the next event, rather then after some delay e, which is close to what we want. The predicate PREFER, defined in [Gal89] is:

(PREFER x P Q) := (BEL x [(BEL x $\Diamond$(P $\vee$ Q)) $\rightarrow$ ((GOAL x $\Diamond$P) $\wedge$ $\neg$(GOAL x $\Diamond$Q))])

Namely, "if x believes that either P or Q will happen, x has the goal that P and not the goal that Q". However, this is a rather crude version of preference, since, as it can be easily verified, one cannot consistently "PREFER" P to Q and Q to K . Gallier's prefer is a strictly binary notion; a graded version of "preference" would require a partial ordering among the $\mathcal{G}$-accessible worlds, which is tantamount to introducing modality.

7. A summary of the properties

Tables A summarizes the relation between C&L's GOAL, P-GOAL and INTEND_1 and a choice of three propositional attitude predicates ("belief", "wish"— as informally defined above— and "intention", as it emerges from the observations in section 5., w.r.t. their closure under material implication (col.I), valid implication (col.II), (currently) believed implication (col.III and IV), and belief that the consequent holds (col.V). P is a state variable, A an action variable. The numbers in the paragraph below correspond to table entries.

Notice that, as recognized by C&L themselves and remarked in [All90], both P-GOAL and INTEND *are* closed under expected implication (contra the third desiderata in section 5., as long as, for (1), Q is adopted as a (fanatical)[22] P-GOAL and the agent believes that Q is currently false. (2) is a "yes" just in case the agent believes that Q is currently false. Therefore, C&L's predicates for "intention" and "goal" do not capture important desiderata for these notions.

8. Intentions with Commitment

Cohen & Levesque's formaliza¿tion of intentions seems to be facing three sources of problems. (1) The lack of a true modal operator. (2) The definition of P-(R-)GOAL. (3) The reduction of intentions to P-GOAL and BEL. Now, problem (3) can be eliminated and problem (2) bypassed if we adopt the theory of intentions that I have discussed above.

First of all, we need to express the notion of the possibility to do an action, using the (traditional) predicate CANDO. We can couch the meaning of CANDO in possible-worlds terms, as follows.

[20] For instance, I might want to be happy *as soon as possible*, meaning with this as soon as the goal of becoming happy does not interfere with the currently conflicting goal of sitting in my office until five. In fact, time-specific goals can be unified with atemporal ones by taking the temporal specification as a special conflicting goal dictating that the desired state should not come about before (or outside) a certain time (or time interval)

[21] Notice that this doesn't mean that the P-GOAL is fulfilled, since I never come to the conclusion that *currently* (BEL me Dead(jb)) or (BEL me $\Box\neg$Dead(jb)). All I can do is sit along the river and wait for his body to pass by.

[22] The problem is solved in this case by the use of P-R-GOAL, where the goal that Q is relativized to the persistence of the agent's belief that (P $\rightarrow$ Q).

TABLE A:	Possibility to derive (*PRED* x Q) given (*PRED* x P) and:				
	$P \rightarrow Q$	$\models P \rightarrow Q$	$(BEL\ x\ (P \rightarrow Q))$	$\Box(BEL\ x\ \Box(P \rightarrow Q))$	$(BEL\ x\ Q)$
Belief	No	No	Yes	Yes	
Wish	No	Yes/?	No	No	No
Intention	No	No/?	No	No	No
$GOAL_{C\&L}$	No	Yes	Yes	Yes	Yes
$P\text{-}GOAL_{C\&L}$	No	Yes	No(Yes)[1]	No(Yes)[2]	No
$INTEND^1_{C\&L}$	No	Yes	No(Yes)[1]	No(Yes)[2]	No

M,σ,v,n $\models$ (CANDO x A) iff for each state P_i which x believes to be a *precondition for A*, M,σ,v,n $\models$ (BEL x P_i) and there is a world σ', $\sigma' \in \mathcal{B}$, such that M,σ',v,n $\models$ (HAPPENS x A),

That is, x believes that all the preconditions for A are true and that there is a possible open situation in which x does the action A as the next thing.[23]

Second, we need a formal definition of *commitment*. For this purpose, we can create L_{BWC}, augmented from L_{BW} with a new accessibility relation T $\times$ P $\times$ Z $\times$ T , $\mathcal{C}$. I will tentatively assume that $\mathcal{C}$ is Eucledian, transitive and serial, and that $\mathcal{C} \subseteq \mathcal{W} \bigcap \mathcal{B}$. Intuitively, $\mathcal{C}$ accesses worlds which the agent regards as both desirable and possible. The worlds in $\mathcal{C}$, then, are those in which the agent does the action(s) that she has decided to do.

Now we can define a predicate COMMITTED, as follows.

M,σ,v,n $\models$ (COMMITTED x A), A an action, iff for all σ^* such that $<\sigma$, n$>$ $\mathcal{C}$[v(x)]σ^*, M,σ^*,v,n $\models$ $\Diamond$(DONE x A).

That is, "x is committed to doing A" in world σ at time n iff it is true in all worlds σ^* accessible to x in world σ at time n via the relation $\mathcal{C}$ the action A *will be accomplished* (this formulation leaves open the possibility that A is currently in execution, as in maintainance actions).[24] Unlike with beliefs, we do not want COMMITTED to be closed under believed implication (i.e. we do not want as an axiom $\forall$x ((COMMITTED x A) $\wedge$ (BEL x((DONE A) $\rightarrow$ (DONE B)))) $\rightarrow$ (COMMITTED x B). The reasons for this will be clear later.

Third, we need a weak, recursive version of "plan-as-recipe", one where we don't necessarily have *all* the steps spelled out from the start. Rather, we believe that either we have the power to bring about the desired state P doing some action W, or we don't; in the latter case, we believe that a complex action is needed, but one which can be segmented in subactions, for each of which we believe we can find a plan.

(D-PLAN W x P) := (BEL x (HAPPENS W;P?)) $\wedge$ (CANDO x W) $\vee$
($\exists K_1 K_2$ (W = K_1;K_2 $\wedge$
(BEL x $\exists J_1 \Diamond$(D-PLAN J_1 x (DONE K_1)))
(BEL x $\exists J_2 \Diamond$(D-PLAN J_2 x (DONE K_2))))))

In the "plan-as-recipe" metaphor, D-PLANs are recipes with very high-level instructions ("make sauce, add to concoction, stir"). Some can be immediately executed ("stir"), while others are complex actions which are currently not fully specified[25].

Now, we can define a new predicate, INTEND, as the composition of COMMITMENT and a D-PLAN.

(INTEND x A) := $\exists$K ((HAPPENS A;K?) $\wedge$ (COMMITTED x A) $\wedge$ (D-PLAN A x K))

How should various feature of intentions follow from these definitions?

(a) *Intentions must be consistent* . This follows from the fact that an agent cannot be committed to two simultaneous conflicting actions, since worlds are internally consistent and, by the definition of COM-

[23] Appealing to "precondition" is little more than handwaving in the present logic system, but it could be made more precise in a framework such as [Lew73], defining "preconditions" as the set of states which must be true in all the accessible possible worlds if the action A is to happen (e.g. $\bigcap$ {P | L((DONE A) $\Box\!\!\rightarrow$ P)} where L is the necessity operator). Notice also that this CANDO is a subjective notion. The rationale is that if agent doesn't know that all preconditions are satisfied he will not try to do A *voluntarily*; on the other hand, if there is some unsatisfied precondition unknown to x, failure to do A will trigger belief revision (this might even be the most economical way to verify if $\mathcal{B}_x$ includes *all* preconditions for A).

[24] In $L_{C\&L}$ actions are a subset of a set E of "primitive event types" [CoL90b,225]. (DONE x A) means, presumably, that a token of an action-type has been executed; however, for an analysis of "commitment" we need to allow a fairly liberal characterizazion of action-tokens, given that an agent might have a commitment to an extremely specific temporal action as well as a commitment to an extremely ill-defined atemporal action, as are those typically required by D-PLAN.

[25] Of course, this notion of plan does not guarantee that the result can be attained. This is how it should be, since an intention to act does not guarantee that the action will be done.

MITTED, the two actions must occur in all the worlds in C.

(b) *Intentions must be realistic* . An agent might believe that she will do some action without having the faintest idea of how to do it. Consider again the case of committed-want; an athlete might say "I want to wint the race!", and believe that she will actually win it. But set aside will-power, she lacks a D-PLAN, i.e. an (at least partial) specification of a procedure to be faster then anybody else in the race. Therefore, she can have a commitment, but not a full-fledged intention.

(c) *Intentions are not closed under expected consequence* . This follows from the fact that the COMMITMENT is *not* closed under (expected) consequence. Notice that D-PLANs by themselves are.[26]

9. Conclusions

In this paper I have presented various aspects of Cohen and Levesque's conversational logic, arguing that their system is not fully satisfactory w.r.t. expressive power and agreement with various philosophical intuitions about goals, beliefs and intentions. I have proposed that the system can be improved within a general theory of the relationship between *motivational* and *cognitive* attitudes that defines *intentions* as a commitment to perform some actions, or assume a certain set of "default belief".

Once modality is brought back into the picture we can give a new definition of INTEND in terms of D-PLANS and COMMITMENT, thus solving three problems associated with the notion of "intention", i.e. their consistency, their realism and the fact that they are not closed under (expected) consequence.

References

[All90] Allen,J. (1990) "A comment on Bratman and Cohen and Levesque." in Cohen, P.R., Morgan, J. and Pollack M.E. *Intentions in Communication*, Cambridge MA, MIT Press.

[Brn84] Brand, M. (1984) "Intending and Acting", Cambridge MA, MIT Press.

[Bra87] Bratman, M.(1987) "Intentions, Plans and Practical Reason", Cambridge MA, Harvard University Press.

[Bra88] Bratman, M., Israel, D.J. and Pollack, M.E. (1988) "Plans and Resource-Bounded Practical Reasoning", Computational Intelligence, vol.4, 4.

[Cas75] Castaneda, H. (1975) "Thinking and Doing", Dordrecht, Holland, D. Reidel.

[CoL90a] Cohen, P.R. and Levesque, H.J. (1990a) "Rational Interaction as the Basis for Communication" in Cohen, P.R., Morgan, J. and Pollack M.E. *Intentions in Communication*, Cambridge MA, MIT Press.

[26] I take that a *plan to A* is also a *plan to C*, if (I know that) A implies C. Under the same conditions, an *intention to A* is not and *intention to C*.

[CoL90b] Cohen, P.R. and Levesque, H.J. (1990b) "Intention is Choice with Commitment", Artificial Intelligence,42, 213-261

[FaH85] Fagin, R., Halpern, J.Y. (1985) "Belief, awareness and limited reasoning: Preliminary Report.", Proceedings IJCAI-85, Los Angeles.

[Gal89] Galliers, J.R. (1989) "A theoretical framework for computer models of cooperative dialogue, acknowledging multi-agent conflict", University of Cambridge Computer Laboratory, Tech.Rep. No.172

[Gef90] Geffner, H. "Default Knowledge: Causal and Conditional Theories" Cambridge, MA; MIT Press.

[Gib90] Gibson, T. "?????" MIT PhD Diss.

[Gri75] Grice, H.P.(1975) "Logic and Conversation" in Cole & Morgan "Syntax and Semantics, 3", New York, Academic Press.

[Haa78] Haack, S. (1978), "Philosophy of Logics", Cambdrige, Cambridge University Press.

[HaM85] Halpern, J.Y., Moses, Y.O. (1985) "A guide to the modal logic of knowledge and belief', Proceedings IJCAI-85, Los Angeles, CA.

[Lew73] Lewis, D. (1973), "Counterfactuals", Cambridge MA, Harvard University Press.

[Per90] Perrault, C.R. (1990) "An application of Default Logic to Speech Act Theory" in Cohen, P.R., Morgan, J. and Pollack M.E. *Intentions in Communication*, Cambridge MA, MIT Press.

[Pol90] Pollack,M. (1990) "Plans as Complex Mental Attitudes" in Cohen, P.R., Morgan, J. and Pollack M.E. *Intentions in Communication*, Cambridge MA, MIT Press.

[ScH90] Schubert, L.K. and Chung Hee, H. (1990) "An Episodic Knowledge Representation for Narrative Texts" University of Rochester, Computer Science, Tech. Rep. 345

[Ven72] Vendler, Z. (1972) "Thoughts" in *Res Cogitans* Ithaca, Cornell University Press.

[WeiIP] Weinberg, A. (In Press) "Minimal Commitment: A Parsing Theory for the Nineties." in *Language and Cognitive Processes.*

[Zam9?] Zamparelli, R. (Forthcoming) "On Generalized Propositional Attitudes", manuscript.